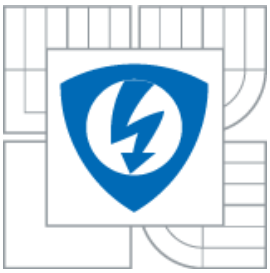
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF BIOMEDICAL ENGINEERING

METODY PREDIKCE AKTIVNÍCH MÍST V PROTEINECH

PREDICTION METHODS OF PROTEIN HOT SPOTS

DIPLOMOVÁ PRÁCE
MASTER´S THESIS

AUTOR PRÁCE — Bc. Jan Duras
AUTHOR

VEDOUCÍ PRÁCE — Ing. Denisa Maděránková
SUPERVISOR

BRNO 2012

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav biomedicínského inženýrství

Diplomová práce

magisterský navazující studijní obor
Biomedicínské inženýrství a bioinformatika

Student:	Bc. Jan Duras	***ID:***	132188
Ročník:	2	***Akademický rok:***	2011/2012

NÁZEV TÉMATU:

Metody predikce aktivních míst v proteinech

POKYNY PRO VYPRACOVÁNÍ:

1) Proveďte literární rešerši používaných metod predikce aktivních míst (hot spots) v proteinech. Zaměřte se především na metody založené na EIIP hodnotách aminokyselin. 2) Pomocí pseudokódu navrhněte realizaci vybrané metody predikce aktivních míst založené na EIIP. 3) Vybranou metodu implementujte v programovém prostředí Matlab. 4) Realizovaný algoritmus validujte na souboru dat a výsledky zhodnoťte srovnáním s publikovanými daty, případně proveďte srovnání získaných výsledků s výsledky z volně dostupných nástrojů pro predikci aktivních míst v proteinech.

DOPORUČENÁ LITERATURA:

[1] FERNANDEZ-RECIO, J. Prediction of protein binding sites and hot spots. WIREs Computational Molecular Science, vol. 1, pp. 680-698, 2011

[2] SAHU, S. S., PANDA, G. Efficient Localization of Hot Spots in Proteins Using a Novel S-Transform Based Filtering Approach, IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, no. 5, pp. 1235-1246, 2011

Termín zadání:	6.2.2012	***Termín odevzdání:***	18.5.2012

Vedoucí práce: Ing. Denisa Maděránková
Konzultanti diplomové práce:

prof. Ing. Ivo Provazník, Ph.D.
Předseda oborové rady

UPOZORNĚNÍ:

ABSTRAKT

Diplomová práce se zabývá metodami vyhledávání aktivních míst v proteinech. V teoretické části je popsáno složení proteinových rozhraní jako i aktivních míst. Jsou zde uvedeny základní principy vyhledávání aktivních míst. V praktické části je vybraná metoda, využívající S-Transformaci, navržena do pseudokódu a následně implementována do programového prostředí Matlab. Program je po té ověřen na souboru vybraných dat a porovnán s dostupnými metodami.

KLÍČOVÁ SLOVA

Protein, aktivní místo, EIIP, rozhraní, interakce, S-Transformace

ABSTRACT

This master's thesis deals with prediction of hot spots in proteins. The theorethical part describes compositions of proteins interfaces and hot spots. There are mentioned basic principles of hot spots prediction. In the practical part is selected method, using the S-Transform, designed to pseudocode and then implmented to Matlab software. Created program is tested on a sample of data and compared with available methods.

KEY WORDS

Protein, hot spots, EIIP, inerface, interaction, S-Transfom

DURAS, J. *Metody predikce aktivních míst v proteinech.* Brno: Vysoké učení technické, Fakulta elektrotechniky a komunikačních technologií, 2012, 28s. Vedoucí diplomové práce Ing. Denisa Maděránková.

Prohlášení

Prohlašuji, že svou diplomovu práci na téma Metody predikce aktivních míst v proteinech jsem vypracoval samostatně pod vedením vedoucího semestrálního projektu a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením tohoto projektu jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č.40/2009 Sb

V Brně dne … .....

podpis autora (autorky)

Poděkování

Děkuji vedoucímu diplomové práce Ing. Denise Maděránkové za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé diplomové práce.

V Brně dne … .....

podpis autora (autorky)

Obsah

Úvod

Motivace pro vyhledávání aktivních míst je velká. Vždyť porozumění biologické funkci proteinů je jedno z nejdůležitějších témat v biologii. Je zde otázka: „Jak primární struktura proteinu definuje konformaci (prostorové uspořádání atomů) a funkci?“. Pokud by se na tuto otázku podařilo odpovědět, začala by nová éra biologie. Ta by umožňovala kontrolu bioaktivity. Tím by byl umožněn například vývoj nových léků působících pomocí malých molekul přímo na rozhraní dvou navzájem reagujících proteinů (ať již inhibičně či excitačně).

Cílem této práce je popsat dostupné metody na určení polohy aktivních míst v proteinech. Existuje mnoho metod pro vyhledávání polohy proteinových vazebních míst (z nich vychází metody na určení polohy aktivních míst), kdy alespoň některé se v této práci snažím představit. Po stručné rešerši se detailně věnuji dvěma metodám založeným na hodnotách EIIP aminokyselin.

V praktické části pak jednu vybranou metodu navrhnu jako pseudokód. A poté ji implementuji do programového prostředí Matlab. Následně provedu její ověření na skupině vybraných dat. A porovnání s ostatními dostupnými metodami a původní metodou.

1. Teoretický úvod – definice pojmů

1.1. Proteiny

Protein je základní stavební kámen všech živých organizmů. Jedná se o biopolymer aminokyselin. Existuje 20 aminokyselin, ze kterých může být protein složen (viz. Tab.1). Typický protein má asi 200 až 300 aminokyselin (menší jsou peptidy). Struktura proteinů je tvořena makromolekulami, které jsou tvořeny posloupností různých aminokyselin v přesně definovaném pořadí. Jejich prostorové uspořádání a biologická funkce je dána složením posloupnosti aminokyselin. Struktura proteinu se dělí na primární, což je lineární pořadí AK. Z této primární struktury lze odvodit strukturu proteinu, mechanismus působení na molekulární úrovni a vzájemné vztahy k jiným proteinům v evoluci. Sekundární struktura nám udává prostorové uspořádání polypeptidového řetězce na krátké vzdálenosti. Tato struktura závisí na složení posloupnosti aminokyselin. Terciární struktura je trojrozměrné uspořádání polypeptidových jednotek (do klubka či vlákna), neboli prostorové uspořádání celého proteinu. Zde dochází k vzájemné interakci postranních řetězců. Kvartérní struktura je skládání polypeptidových jednotek (ne všechny proteiny mají kvartérní soustavu).[1]

Funkční úloha proteinu může být buď dynamická jako transport, kontrola metabolismu, kontrakce nebo katalýza chemických přeměn. Nebo strukturální jako výstavba orgánů, tkání a jejich podpůrné funkce. Dále můžeme proteiny rozdělit podle jejich biologické funkce (enzymy, zásobí proteiny, hormony, …).[1]

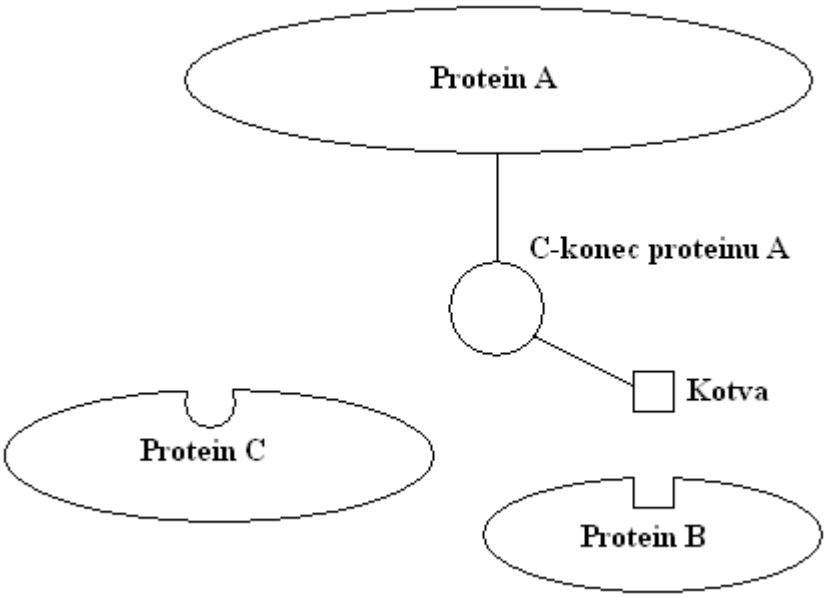
Tab.1: Seznam aminokyselin a jejich zkratek, [2]

Název	**3 písmenná zkratka**	**1 písmenná zkratka**
Alanin	Ala	A
Arginin	Arg	R
Asparganin	Asn	N
Asparágová kyselina	Asp	D
Cystein	Cys	C
Glutamin	Gln	Q
Glutamová kyselina	Glu	E
Glycin	Gly	G
Histidin	His	H
Leucin	Leu	L
Isoleucin	Ile	I
Lysin	Lys	K
Methonin	Met	M
Fenylalanin	Phe	F
Prolin	Pro	P

Serin	Ser	S
Threonin	Thr	T
Tryptofan	Trp	W
Tyrosin	Tyr	Y
Valin	Val	V

1.2. Vazebná místa

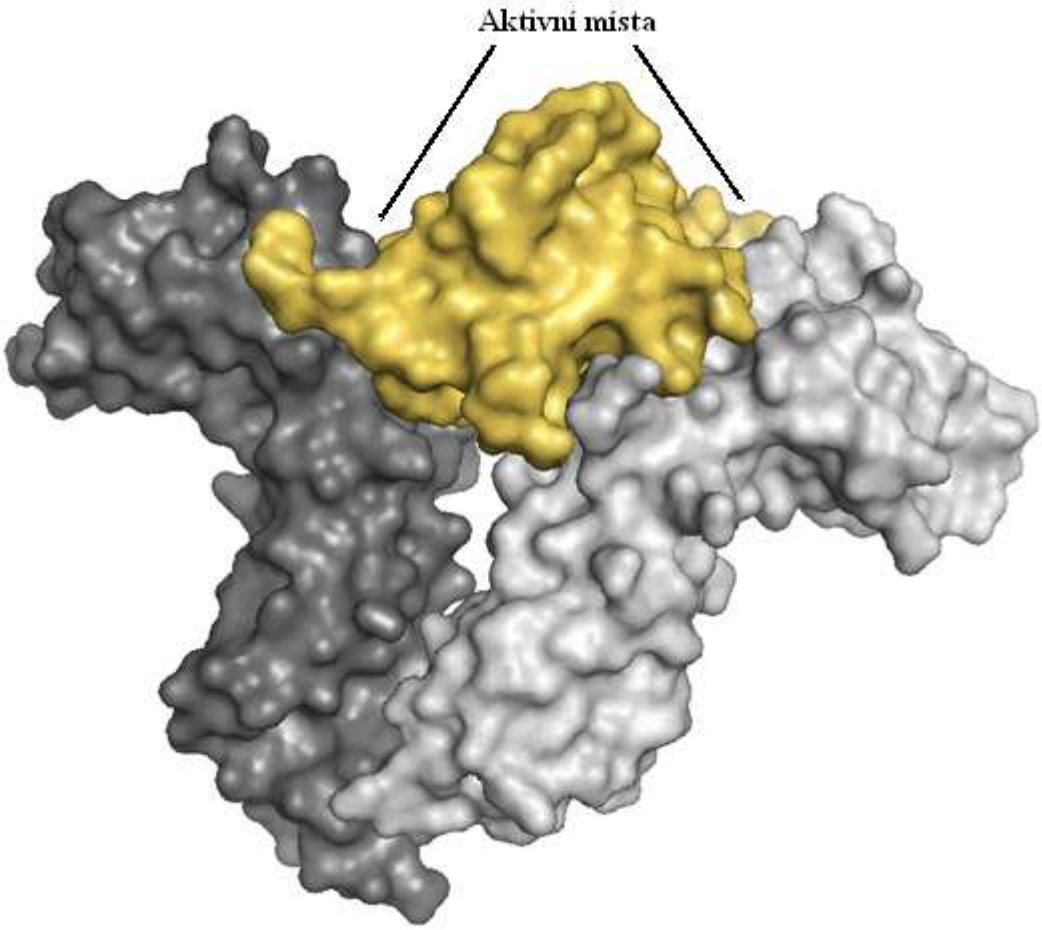
Vazebné místo proteinů je místo, kde spolu dva proteiny reagují. Tyto vzájemné interakce jsou základním kamenem skoro všech procesů v biologii. Je tedy velmi důležité jim detailně porozumět. Právě pro toto porozumění je velmi důležitá analýza vazebních interakcí mezi proteiny a obzvláště vazebních míst, kde dochází k jejich skutečné vazbě na sebe. Tato místa se označují jako proteinová vazebná místa, ta obzvláště významná pro tvorbu proteinového komplexu se nazývají aktivní místa. Okolí vazebného místa proteinu je obvykle ploché a velké ve srovnání přímo s jednotlivými aminokyselinami zodpovědných za vazbu proteinů. Chemické složení vazebních míst a fyzikálněchemické vlastnosti jsou velmi různorodé. Tvar a složení velmi ovlivňují typ interakce, ke které mezi proteiny dochází. Pro homodimery (dimer – molekula složená ze dvou menších polypeptidových řetězců, homodimer – složen z identických menších polypeptidových řetězců) je celkem snadné určit jejich vazební místa. Díky jejich odlišným fyzikálněchemickým vlastnostem vůči místu kde nedochází k interakci. Interakce proteinů je však u heterodimerů (heterodimer – složen z různých menších polypeptidových řetězců) o mnoho složitější. Nedají se tedy tak snadno poznat od zbytku povrchu. Struktura rozhraní dvou reagujících proteinů se dělí na jádro a okolí vazby. Jádro zabírá 2/3 kontaktního povrchu. Okolí vazby je podobné exponovanému povrchu. Pro molekulární rozpoznávání slouží tzv. „kotva“. Ta je „tužší“ (hůře se formuje) než zbytek rozhraní a dobře koreluje se „zachovanými“ aktivními místy (tedy s aktivními místy neochotnými mutovat). Tato „kotva“ je navázána na C-koncové aminokyselině proteinu (konec řetězce aminokyselin proteinu zakončeného volnou karboxylovou skupinou [-COOH]). Velmi často se jedná o komplex, který se nazývá glykosylfosfatidylinositol (GPI). Navíc se ukázala i korelace rozhraní s typem, velikostí a povahou komplexu. Při snaze porozumět proteinovým interakcím nelze opomínat allosterické vlivy. U allosterických proteinů se další protein (např. inhibiční faktor) váže jinam než na aktivní místo. V důsledku toho dojde ke změně konformace proteinu a tím i ke změně aktivního místa. [4], [11]



Obr. 1: Schématické znázornění funkce kotvy u vazeb proteinů. Protein A je schopen pomocí kotvy rozpoznat kompatibilní protein B, protein C nikoli. Ten není kompatibilní s kotvou proteinu A

1.3. Aktivní místa

Aktivní místa proteinů jsou ta vazebná místa s většinou vazebné energie. Byly objeveny, když se zjistilo, že při mutaci určitých reziduí se vazební energie celého proteinu výrazně sníží. Pokles vazebné energie znamená, že protein má menší snahu se vázat. Tím pádem místa u kterých při jejich mutaci výrazně poklesne vazebná energie jsou extrémně důležitá pro vznik vazby mezi proteiny. Termín „aktivní místo“ byl poprvé požit k popisu klíčových reziduí v komplexu lidského růstového hormonu, kde se postupně mutovala vazebná místa na alanin a měřil se přínos jednotlivých postraních řetězců (vazebních míst) na volnou vazebnou energii celého proteinu (ΔG). Takto se identifikovaly 2 rezidua tryptofanu, která poskytovala většinu změny volné vazebné energie ($\Delta\Delta G$). Konkrétně u těchto dvou bylo $\Delta\Delta G$ při jejich mutaci více než 4,5 kcal.mol^{-1} . Časem se stanovila hranice pro určení aktivního místa na rozmezí 1-2 kcal.mol^{-1} změny volné vazebné energie ($\Delta\Delta G$). Aktivní místa jsou velmi „konzervativní“(nenáchylná k mutacím) a jsou obvykle obklopena poměrně „konzervativními“ vazebními místy. Ta spolu tvoří vysoce kooperativní interakce. Strukturně se nachází v okolí center vazebních míst a jsou chráněna před hromadným rozpouštěním energeticky méně významnými vazebními místy. Ta tvoří hydrofobní *O-ring* (ten chrání aktivní místa před rozpouštědly). Nejčastěji aktivní místa tvoří tryptofan, arginin a tyroxin, zatímco leucin, serin, theronin a valin jsou málo obvyklé. [4]



Obr. 2: Schématický pohled na komplex lidského růstového hormonu a jeho receptoru. Lidský růstový hormon (žlutý) se váže na extracelulární homodimerické receptory (šedivé),[2]

2. Metody predikce vazebních míst proteinů

2.1. Metody v laboratořích

Metody v laboratořích jsou velmi drahé, zároveň ale důležité pro porovnávání výsledků výpočetních metod. Mezi těmito metodami je nejdůležitější metoda *Alanin Scanning Mutagenesis* (dále jen ASM). U této metody se postupně nechají všechny aminokyseliny zmutovat na alanin a měří se jejich energetický příspěvek (změna volné vazebné energie proteinu ΔΔG) ke schopnosti interakce proteinu. Tato metoda je trochu nejednoznačná, protože každá jednotlivá mutace nemůže popisovat snahu (výkon) interakce celého proteinu. Jeho interakce jsou velmi komplexní, aby byly počítány jako důsledky mutací jednotlivých aminokyselin. Nicméně je tato metoda akceptována a používá se pro porovnávání výsledků výpočetních metod. Její hlavní nevýhodou je časová a finanční náročnost.[3]

2.2. Výpočetní metody pro predikci vazebních míst proteinů

Pro určení vazebních míst proteinů pomocí výpočetních metod jsou dva základní přístupy. Na základě sekvencí, nebo na základě prostorově strukturních informací. V současnosti se používají oba směry. Je ale snahou vyvíjet metody na základě sekvencí z důvodů nepotřebnosti draze získávaných informací o prostorové struktuře proteinů. Informace o sekvenci proteinu (jeho posloupnosti aminokyselin) se získávají levněji. Ale u obou přístupů platí, že úspěšnost metody velmi záleží na testovacích datech. Často se stává, že metoda na daném testovaném vzorku má slušné výsledky. Ale po změně testovacích dat nastává pokles prediktivních schopnosti i přesnosti metody.

2.2.1. Metody predikce vazebních míst proteinů na základě prostorové struktury

V současnosti stále tyto metody převažují. Využívají informace o prostorové struktuře proteinu, popřípadě proteinových komplexů. Jsou často kombinovány s dalšími fyzikálně-chemickými ale i jinými parametry. Mají stále lepší prediktivní výsledky než metody čistě založené na základě sekvenčních informací. Ty ale na druhou stranu mohou být užity ve více případech (chybějící informace o struktuře proteinu, resp. proteinových komplexů). [4]

Empirická skórovací funkce

Tyto metody jsou založené na získaných fyzikálněchemických parametrech nebo strukturálních vlastnostech z povrchu proteinového komplexu. To je možné jen u určitých typů interakcí, resp. jen u některých lze získat určitá specifická data která je odlišují od zbytku proteinu. Z těchto informací jako např. skon rozhraní, konzervační skóre, solvatační potenciál, hydrofobnost se snaží vytvořit funkce pomocí kterých, se pokouší predikovat vazebná místa. Mezi tyto metody patří *Inter-ProSurf* (viz. Tab. 2), *SiteEngines* (viz. Tab. 2), ta navíc využívá hierarchické vyhodnocování a *Protein IntErface Recognition* (viz. Tab. 2). [4]

Zachování sekvence

Tyto metody využívají zachování sekvence nebo evoluční informace. Vycházejí z předpokladu, že rozhraní nebudou v čase příliš ochotně mutovat z důvodu zachování biologické funkce proteinu, resp. komplexu. Tyto metody využívající zachování sekvence, kombinují tuto vlastnost s dalšími parametry např. s fyzikálními a empirickými parametry. Mezi tyto metody patří *Evolutionary Trace* (viz. Tab.2), *ConSurf* (viz. Tab.2), *Joint Evolutionary Trees* (viz. tab.2), podobná *TreeDet server* (viz. tab.2), *Promate server* (viz. Tab.2), *pyDockRST* metoda (ta dosáhla na určitých specifických datech excelentních výsledků), *PRotein-protein Interaction prediction by Structural Matching* (viz. Tab.2), *Protein Interface residUe Prediction* (viz. tab.2) tato metoda dosáhla při optimalizaci na určitá specifická data pozitivní prediktivní hodnoty (dále jen PPH) 44,5% a sensitivity 42,4%, ale pro jiná data nastal pokles, konkrétně PPH na 29,4% a sensitivita na 30,5%.[4]

Učící se techniky

U těchto metod se využívá různých učících se technik a klasifikátorů. Často se využívá neuronová síť a pomocný mechanický vektor. Častá je i jejich kombinace. Tyto metody se pomocí zmíněných klasifikátorů snaží roztřídit protein (na menší části, čímž přiblíží určení vazebního místa) tak aby nejlépe oddělil vazebná místa a zbytek. K tomu se využívá různých vlastností proteinů K těmto metodám patří *cons-Protein-Protein Interaction Site Prediction* (viz. Tab.2), *Patch Finder Plus* (viz. Tab.2), *Protein-Protein Interface PREDiction* (viz. Tab.2) a *Protein-Protein Interface iDEntification and Recognition* (viz. tab.2). Dá se obecně říci, že mají dobrou PPH (mezi 50 - 70%), ale horší sensitivitu (okolo 20%).[4]

Meta Servery

Tyto metody jsou servery pracující nejčastěji s kombinací neuronové sítě a další metodou. Tím zvyšují účinnost původní metody. Hlavní výhoda těchto serverů spočívá v pohodlném přístupu k více metodám. Ale při využití meta serverů, bychom měli být opatrní na interpretaci výsledků. Měl by se zhodnotit přínos jednotlivých metod. Známý je hlavně server *meta-PPISP* (viz. Tab.2), který kombinuje neuronovou sít s metodou *cons-Protein-Protein Interaction Site Prediction*. [4]

Energetické metody

Tato skupina metod využívá hlavně energetické parametry jako jsou solvatační potenciál, hydorfobnost atd. Významným zástupcem této kategorie metod je *Optimal Docking Area* (dále jen ODA) (viz. Tab.2). Je založen na hypotéze, že desolvatace (desolvatace – oddělení částic rozpouštěné látky od částic rozpouštědla) má ústřední význam pro vznik proteinové vazby. Konkrétně tato metoda využívá počítačového algoritmu na určení souvislého povrchu s optimální dokovací energií. Dokování je metoda, která pro dvě různé molekuly najde uspořádání v jakém mohou tyto dvě molekuly intereagovat. Pokud takovéto uspořádání existuje najde nejvhodnější orientaci obou molekul z hlediska největší komplementarity a z hlediska maximálního počtu nevazebných interakcí a nejmenší energie celého komplexu. Velikost výpočetní plochy není pevná. Počítá se, dokud není nalezena

kruhová plocha s nejpříznivější desolvatační energií z každého počátečního bodu (počáteční body jsou definovány pomocí parametru rozpustné přístupové plochy nebo v centrech reziduí vazebních míst postranních řetězců). ODA dosáhla na jistém testovacím vzorku úspěšnosti 80%, ale je použitelná pouze tehdy, kdy je doslvatační efekt významný (asi ½ případů). [4], [10]

2.2.2. Metody predikce vazebních míst proteinů na základě sekvence

Zatím existuje málo metod založených pouze na informaci o sekvenci. Tyto metody využívají hlavně dva přístupy. Za prvé je možné použít informace z korelace mutací z více zarovnaných sekvencí proteinů. Tento přístup vychází z hypotézy, že residua zapojená do „kontaktů“ proteinů mají tendenci mutovat současně během evoluce. Druhý přístup je založený na analýze distribuce hydrofobnosti podél sekvence. U tohoto přístupu je senzitivita mezi 59% a 80% v závislosti na metodě a užitém souboru dat. [4]

I u metod založených na sekvenci se využívají různé učící se přístupy a klasifikátory, především pomocný mechanický vektor (tyto metody mají podobnou senzitivitu, ale spíše nižší PPH , jako přístup založený na distribuci hydrofobnosti). *Interaction Sites Identifed from Sequence* (ISIS)(viz. Tab.2) je také metoda využívající učící se postup, který kombinuje s evoluční informaci a predikcí strukturálních vlastností. Predikce dosáhly velmi vysoké PPH, ale za cenu malé senzitivity. To znamená, že metoda předpokládá velmi málo zbytků, ale s vysokou přesností. Z toho vyplývá, že by se zde mohla nacházet důležitá rozhraní. Tím se dostáváme k problematice vyhledávání aktivních míst v proteinech.[4]

Tab. 2: Přehled metod pro predikci proteinových vazebních míst,[4]

Název metody	**Vstupní data**	**Metodologie**	**Detaily**
ISIS	Sekvence	Neuronová síť	predikuje strukturální vlastnosti, evoluční informace
TreeDet	Sekvence, struktura	Skórovací funkce	sekvenční a strukturální zarovnání
Promate	Struktura	Skórovací funkce	sekundární struktura, zachování sekvence, typ residua
PINUP	Struktura	Skórovací funkce	skóre energie postranního řetězce, sklon, zachování sekvence
InterProSurf	Struktura	Skórovací funkce	rozpustná přístupnost, skon
PRISM	Struktura	Skórovací funkce	geometrické doplňky, zachování
ConSurf	Struktura	Skórovací funkce	Zachování
ET	Struktura	Skórovací funkce	mnohonásobné zarovnání sekvence
JET	Struktura	Skórovací funkce	strukturální a funkční zachování
WHISCY	Struktura	Skórovací funkce	zachování, povrchové vlastnosti
PIER	Struktura	Skórovací funkce	atomární statistické vlastnosti
SiteEngines	Struktura	hierarchické skořovací Funkce	strukturální shoda, fyz.chem. vlastnosti
PPI-Pred	Struktura	Neuronová síť	tvar povrchu, elektrostatický pot.
cons-PPISP	Struktura	Neuronová síť	PSI-Blast sekvenční profil, rozpustná přístupnost
SPPIDER	Struktura	Neuronová síť	rozpustná přístupnost a další vlastnosti
Patch Finder Plus	Struktura	Neuronová síť	zachování, vydutost, plocha, vodíkové vazby, frekvence residuí
meta-PPISP	Struktura	Meta server	cons-PPISP, Promate a PINUP
PI^2PE	Struktura	Meta server	cons-PPISP, WESA, DISPLAR
SHARP	Struktura	Energeticky založená,	sklon, desolvatace, hydrfobnost, ASA,

Tab. 2: Přehled metod pro predikci proteinových vazebních míst,[4] (pokračování)

SHARP	Struktura	skórovací funkce	tvar povrchu
ODA	Struktura	Energeticky založená	desolvatační energie
NIP	Struktura	Energeticky založená	dokovací simulace

3. Metody predikce aktivních míst

Existují různé přístupy jak predikovat aktivní místa. Vyvinuly se různé bodovací systémy na základě zachování reziduí, vazby vodíku, vazební energie popřípadě její změně. Další přístupy na základě kombinací výše uvedených parametrů v kombinaci se strojově učícími metodami. Většina metod je založena na základě struktury proteinů a málo je jich jen na základě sekvencí.[4]

Aktivní místa jsou velmi „konzervativní“ (neochotná mutovat) a jsou obklopena středně „konzervativními“ a energeticky méně významnými rezidui. Ty tvoří hydrofobní *O-ring*. Ten chrání aktivní místa před rozpouštědly. Zdá se, že jsou seskupeny v těsných balíčcích v centrech rozhraní dvou proteinů, ale zatím nebyl zjištěn jediný atribut definující aktivní místo sám o sobě (jako tvar, náboj, hydrofobicita,…). [4]

3.1. Predikce aktivních míst na základě prostorové struktury

Většina těchto metod je založena na informací ze 3D struktury proteinu. Využívají se empirické metody (velmi často se vyskytují skórovací data, založená na zachování sekvence), také se využívají metody založené na energetických úvahách. [4]

Empirické metody

V minulosti se prosazoval přístup založený na vyhledávání charakteristických prvků. V současnosti je pohled na rozhraní proteinů takový, že rozhraní je tvořeno z různých vazebných míst. Ta jsou zapojena do specifických interakcí spolu se skupinou „zachovaných“ aktivních míst. Ty působí jako vazebné místo pro „kotvu“ (důležitá pro proteinové rozpoznávání). Se zvyšující se velikostí interakce roste i počet aktivních míst. [4]

Vzhledem k energetické důležitosti aktivních míst se očekává „konverzace“ aktivních míst na rozhraních podle dané rodiny proteinů. Tento fakt využívá metoda *Multiple Aligment of Protein-Protein InterfaceS* (MAPPIS) (viz. Tab.3). MAPPIS úspěšně předpovídá aktivní místa. Poskytuje celkem dobrou PPH, ale je zapotřebí dostatečného počtu proteinových komplexů ve vysokém rozlišení z funkčně podobných proteinů. [4]

Energeticky založené metody

Několik metod je založeno na výpočetním alaninovém skenování proteinového komplexu. To spočívá ve výpočtu variace vazební afinity ($\Delta\Delta G$) při počítačovém mutování jednotlivých vazebních míst na alanin. Tento přístup využívá metoda *ROBETTA* (viz. Tab.3). Ta dosahuje vysoké PPH na určitých datech, ale na jiných datech není již tak úspěšná. Mezi další metody patří *FOLD-X Energy Function* (FOLDEF) (viz. Tab.3). Ta dosahuje 61% PPH a 72% sensitivity. U větších datasetů sice mírně vzroste PPH, ale významně klesne sensitivita.[4]

Predikce aktivních míst založené na nevázané proteinové struktuře

Výše uvedené metody pracující na základě proteinového komplexu mají limitaci v potřebě 3D struktury komplexu nebo úzce homologních proteinů. Většina proteinových

interakcí však není ve 3D struktuře dostupná a z toho vyplývá i omezení uvedených metod. Vyvíjejí se proto predikce aktivních míst na základě proteinové dokovací výpočetní simulace. Ty jsou vhodné při nepřítomnosti znalosti 3D struktury. Do těchto metod patří metoda *pyDockNIP* (viz. Tab.3) Tato metoda měla na testovacím datasetu PPH 68% a sensitivitu 43%.[4]

3.2. Predikce aktivních míst na základě sekvence

Jak je uvedeno výše, není mnoho metod pro predikci aktivních míst čistě na základě sekvencí. V této práci jsou uvedeny tři, z toho dvě jsou popsány podrobněji (dále).

Metoda ISIS (viz. Tab.3) (zmíněna v 2.2.2) byla původně navržena pro predikci proteinových rozhraní, ale následně byla upravena pro predikci aktivních míst. Byla testována na datasetu 296 mutací z 30 různých komplexů. Dosáhla pozoruhodně vysoké prediktivní míry. To lze vysvětlit použitím úzce vymezeného měřítka pro predikci. To zavedlo aktivní místo a bylo označeno pokud $\Delta\Delta G > 2{,}5 kcal.mol^{-1}$ a místa, kde nejsou aktivní místa pouze pro $\Delta\Delta G=0$ $kcal.mol^{-1}$, proto se vyřadily všechny mutace s $\Delta\Delta G<0$ ač měly být označeny jako místo kde není aktivní místo. Také mutace s $0<\Delta\Delta G<2{,}5 kcal.mol^{-1}$,kde by potenciálně mohla být aktivní místa jsou vyřazena jako místa kde nejsou aktivní místa.[4]

Tab.3: Přehled metod pro predikci aktivních míst,[4]

Název Metody	**Vstupní data**	**Metodologie**	**Detaily**	**Sesitivita**	**PPH**
ISIS	Sekvence	Neuronová sít	predikuje strukturální vlastnosti, evoluční informace	15%	89%
FOLDEF	Struktura komplexu	Energeticky založená	Alaninové skenování	45-72%	61-73%
ROBETTA	Struktura komplexu	Energeticky založená	Alaninové skenování	28-69%	60-71%
K-FADE/K-CON/ ROBETTA	Struktura komplexu	Mechanicky učící se algoritmus	fyzikální a biochemické vlastnosti	48%	53%
MAPPIS	Struktura komplexu	Evoluční zachování	mnohonásobné zarovnání, 3D shlukování	66%	63%
HotPoint	Struktura komplexu	Empirický model	přístupnost, potenciály vycházející ze zkušeností	59%	70%
pyDockNIP	Struktura nevázaného proteinu	Energeticky založená	dokovací simulace	42-43%	68-75%

3.3. Predikce aktivních míst na základě převodu sekvence na digitální signál

Tyto metody využívají převodu sekvence nukleotidů proteinu na aminokyseliny a poté na číselné sekvence. Tyto sekvence se získají pomocí převodu vycházejícího z různých fyzikálněchemických a biologických vlastností. Existuje více možností, ale mezi nejvíce používané patří *Ionization Constant* (dále jen IC), která vyjadřuje kyselinovou disociační konstantu z příslušné ionizační reakce a především *Electon-Ion Ineraction Pseudo-potencial* (dále jen EIIP).

Hodnoty **EIIP (Electron – Ion Interaction Potential)** – česky: „elektronový interakční potenciál“, vyjadřuje průměrný energetický stav valenčních elektronů. Neboli distribuci volných elektronů podél sekvence aminokyselin. Pokud se tyto hodnoty použijí k převodu sekvence aminokyselin na diskrétní signál vyjadřuje potom výsledný signál tyto vlastnosti pro celou sekvenci. Při určení hodnot EIIP pro jednotlivé aminokyseliny se vychází z hodnot EIIP pro jednotlivé nukleotidy, ze kterých jsou složeny aminokyseliny (viz. Tab.4).

Tab. 4: Hodnoty EIIP pro nukleotidy

Adenin (A)	0,1260
Thymin (T)	0,1335
Guanin (G)	0,0806
Cytosin (C)	0,1340

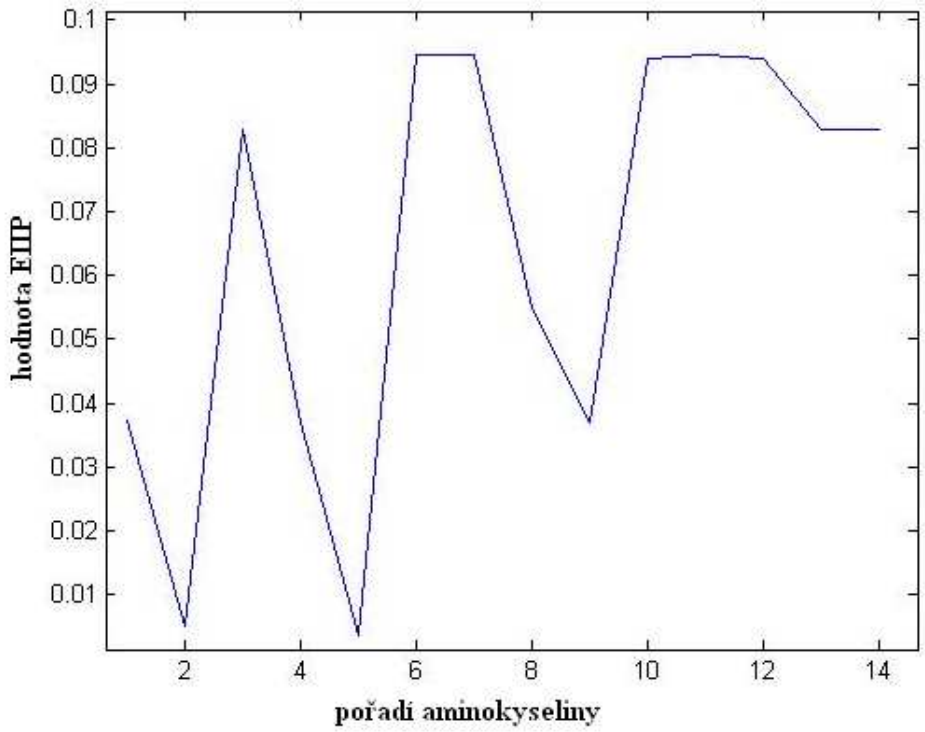
Z těchto hodnot se vyšlo při určování hodnot EIIP pro jednotlivé aminokyseliny, ty jsou ukázány v tabulce 5.

Tab. 5: Hodnoty EIIP pro jednotlivé aminokyseliny

Název aminokyseliny	**3 písmenná zkratka**	**1 písmenná zkratka**	**EIIP**
Leucin	Leu	L	0,0000
Isoleucin	Ile	I	0,0000
Asparganin	Asn	N	0,0036
Glycin	Gly	G	0,0050
Valin	Val	V	0,0057
Glutamová kyselina	Glu	E	0,0058
Prolin	Pro	P	0,0198
Histidin	His	H	0,0242
Lysin	Lys	K	0,0371
Alanin	Ala	A	0,0373
Tyrosin	Tyr	Y	0,0516
Tryptofan	Trp	W	0,0548
Glutamin	Gln	Q	0,0761
Methonin	Met	M	0,0823
Serin	Ser	S	0,0829
Cystein	Cys	C	0,0829
Threonin	Thr	T	0,0941

Fenylalanin	Phe	F	0,0946
Arginin	Arg	R	0,0959
Asparágová kyselina	Asp	D	0,1263

Na obr. 3 Je vidět příklad převodu sekvence aminokyselin na diskrétní signál.



Obr. 3: Příklad převodu sekvence aminokyselin na diskrétní signál. Zde se jedná o somatostatin (inhibiční faktor lidského růstového hormonu) s původní sekvencí aminokyselin: Ala-Gly-Cys-Lys-Asn-Phe-Phe-Trp-Lys-Thr-Phe-Thr-Ser-Cys.

3.3.1. Metoda využívající frekvenční přístup

Jedná se o metodu uveřejněnou v [5]. Jejím základem je převod sekvence na digitální signál. Pro převod využívá hodnot EIIP i IC (*Ionisation Constant* – Ionizační Konstanta) a jeho následné digitální zpracování. Poté využívá klasifikaci pomocí učících nástrojů. Pracuje také s počítačovým alaninovým skenováním. Hodnoty IC v tab. 6, hodnoty EIIP v tab.5.

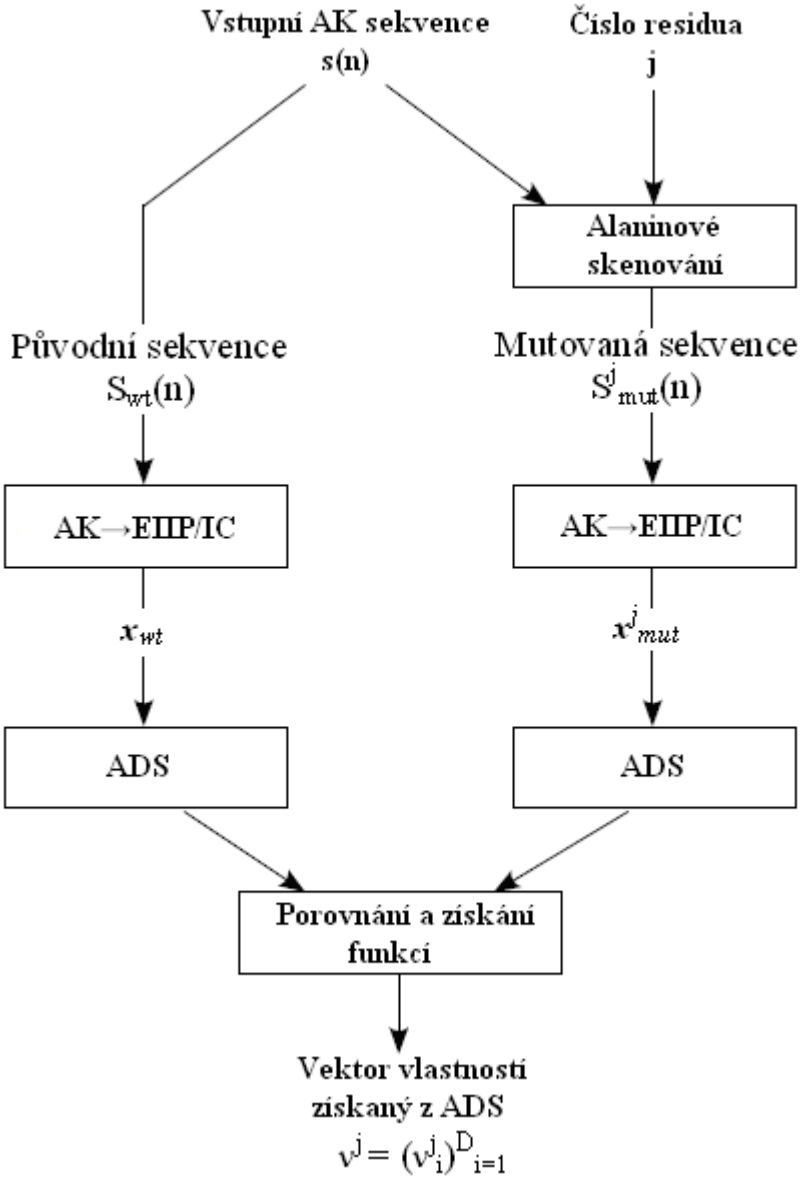
Tab. 6: Hodnoty EIIP pro jednotlivé aminokyseliny

Název aminokyseliny	**IC**
Leucin	2,4000
Isoleucin	2,4000
Asparganin	2,2000
Glycin	2,4600
Valin	2,3500
Glutamová kyselina	2,3000

Prolin	2,0000
Histidin	2,3000
Lysin	2,2000
Alanin	2,3000
Tyrosin	2,2000
Tryptofan	2,3700
Glutamin	2,0600
Methonin	2,1700
Serin	2,1000
Cystein	1,9600
Threonin	2,0900
Fenylalanin	1,9800
Arginin	1,8200
Asparágová kyselina	1,8800

Počítačové alaninové skenování je užitečné v analýze rozhraní proteinů. Zbytky vazebných míst se zmutují na alanin (tj. odstranění postranního řetězce za určitým uhlíkovým atomem) a vyhodnocuje se vliv této mutace na vazební afinitu proteinového rozhraní. Tedy měří se volná vazebná energie ($\Delta\Delta G$). Metoda je čistě výpočetní a je inspirována experimentální ASM(zmíněna v 2.1). Postup je podobný jako u ASM, ale vše se provádí výpočetně, simulací na počítači. Nahrazují se subsekvence postraních řetězců alaninem a hledají se frekvenčně spojené změny v celé sekvenci. Postup je podobný jako v [6], ale místo zkoumání fyzického modelu nebo jednoduchého měření volné vazebné energie se zde analyzují změny ve frekvenčním spektru způsobené mutací. [5]

Postup metody je následující. Nejdříve se provede mutace okolo zvolené pozice pomocí výpočetního alaninového skenování. Místo výměny pouze zbytku vazebného místa s(j) je zde užito okno s centrem v pozici j. Pozice j je postupně posouvána po celé sekvenci. To je celé zpracováno, tedy celá sekvence v okně bude zmutována na alanin. Změna jednoho vzorku neměla výrazný vliv na celkové spektrum, byť je to v rozporu s teorií *O-ringu*. Okno je zde voleno délky 5(délka byla zjištěna empiricky).Tato délka okna navíc respektuje případy, kdy jsou aktivní místa blízko vedle sebe. Po mutaci jsou obě sekvence, původní $s_{wt}(n)$ a mutovaná $s^{j}_{mut}(n)$ mutacemi v okolí místa j, převedeny pomocí hodnot EIIP na diskrétní signál. Jsou analyzovány stejnými nástroji pro analýzu digitálního signálu (dále jen ADS). Celý proces je znázorněn na obr.4. Poté jsou porovnávány vlastnosti vypočítané z frekvenčního spektra obou signálů (resp. sekvencí). Zde konkrétně je užito srovnání změn vrcholů spektra, změny energie na dílčích částech sekvence a změna globální energie. [5]



Obr. 4: Schéma metody založené na frekvenčním přístupu,[5]

Změna vrcholů spekter je počítána jako poměr lokálních maxim spekter (1) původní a mutované sekvence.

$$I = \left\{0<k<N : \left|X_{wt}(k)\right| \geq \max(\left|X_{wt}(k-1)\right|, \left|X_{wt}(k+1)\right|)\right\}. \qquad (1)$$

Kde $\boldsymbol{X_{wt}}$ = FFT ($\boldsymbol{x_{wt}}$) je FFT z $\boldsymbol{x_{wt}}$ a FFT má rozměr $\boldsymbol{N}$, $\boldsymbol{N}$ je rovné délce sekvence a $\boldsymbol{I}$(1) nám dá pozice lokálních maxim. FFT znamená rychlá fourierova transformace (*Fast Fourier Transfomation*). Stejnosměrná složka (vzniká v důsledku shodnosti některých hodnot EIIP) je odstraněna ještě před FFT. Potom se pro body, kde je lokální maximum (vrchol spektra), provede výpočet samotných změn vrcholů spekter.

$$ZmenaVrcholu_k^j = \frac{|X_{wt}(k)|}{|X_{mut}^j(k)|}. \quad (2)$$

Kde $\boldsymbol{X^j_{mut}}$ = FFT ($\boldsymbol{x^j_{mut}}$) a k patří k posuzované frekvenci. V této metodě se vyberou pouze tři největší změny jako deskriptory. Díky symetrii spekter mohu brát v úvahu pouze první polovinu jednotlivých spekter.[5]

Změna energie v dílčích částech sekvence je změna lokální energie ve frekvenčních dílčích pásmech. Sekvence se transformuje do časově frekvenční reprezentace pomocí STFT s posuvným oknem délky $\left(\frac{N}{4}+1\right)$, kde $\boldsymbol{N}$ je rovno délce sekvence. STFT znamená fourierova transformace s krátkým časem (*Short Time Fourier Tranformation*) Analyzované okno má malé postranní laloky,proto se požije se 4 výrazové Blackman-Harrisovo okno(na obr.5, je zevšeobecněním rodiny Hammingových oken s minimalizací vedlejších laloků). To bylo přijato jako kompromis mezi šířkou hlavního laloku a úrovní postraních laloků. Po STFT díky symetrii frekvenčních spekter $\boldsymbol{S^j_{mut}(j,.)}$ a $\boldsymbol{S_{wt}(j,.)}$ můžeme vzít pouze jejich dolní polovinu. Ta je poté rovnoměrně rozdělena do osmy stejných dílčích pásem. Změna energie způsobená výpočetním alaninovým skenováním (mutací) bude pozorována v těchto pásmech. Pomocí výpočtů:

$$SBZmenEnergie_m^j = \frac{\sum_{\nu \in SBm} |S_{wt}(j,\nu)|^2}{\sum_{\nu \in SBm} |S_{mut}^j(j,\nu)|^2}, m = 1...8, \quad (3)$$

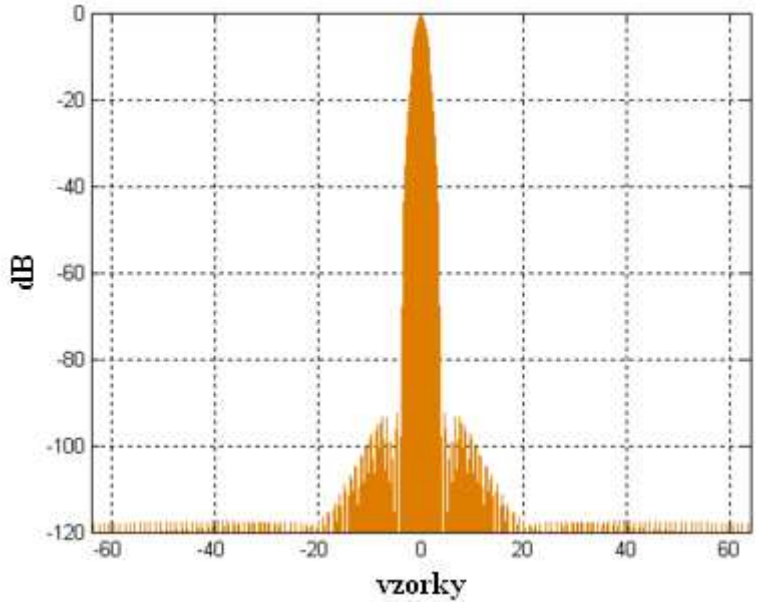
kde $\boldsymbol{S_{wt}}$ = STFT($\boldsymbol{x_{wt}}$) a $\boldsymbol{S^J_{mut}}$ = STFT($\boldsymbol{x^j_{mut}}$) a $\boldsymbol{SBm}$(4) je m-té dílčí pásmo.

$$SBm = \left\{ k : (m-1)\frac{N}{16} \le k \le m\frac{N}{16} \right\} \quad (4)$$

Poslední počítaná vlastnost je změna globální (celkové) energie. Jde o poměr energie mutované sekvence vůči energii původní sekvence. [5]

$$ZmenaEnergie^j = \frac{\sum_{n=1}^{L} |x_{mut}^j(n)|^2}{\sum_{n=1}^{L} |x_{wt}(n)|^2}. \quad (5)$$

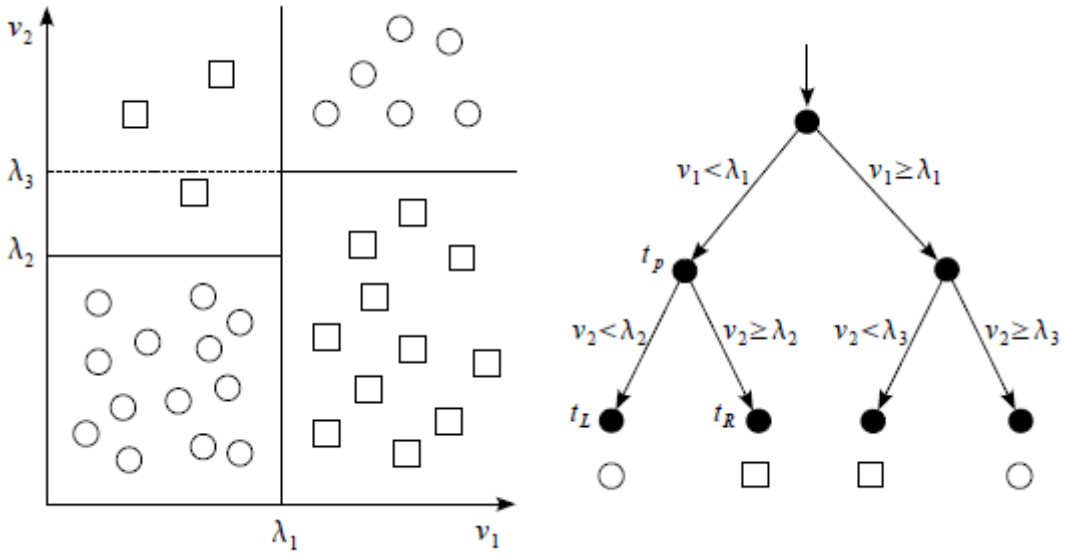
Kde $|\boldsymbol{x^j_{nut}(n)}|^2$ je energie mutované sekvence a $|\boldsymbol{x_{wt}(n)}|^2$ je energie původní sekvence. A $\boldsymbol{L}$ je délka sekvence. [5]



Obr. 5: Frekvenční reprezentace 4 výrazového Blackman-Harrisova okna,[7]

Pro zhodnocení navržených deskriptorů potenciálně sloužících pro identifikaci aktivních míst se využívá učící klasifikátor. Zde je zvolen Random Forest (*Random Forest* - náhodný les, dále jen RF), což je jedna z nejvýkonnějších metod pro zařazování pod dohledem. RF se skládá z *Klasifikačních Stromů* (dále jen KS). Ty jsou základním komponentem RF. KS je strukturovaný prediktivní model, ve kterém každý vnitřní uzel je spojen s rozhodovacím pravidlem. Ta jsou založena na vlastnostech objektu $\boldsymbol{v = (v_i)_{i=1}^{i=D} \in V}$, kde V je funkcí prostoru. Každý konečný list (koncový bod) je přidělen do třídy $\boldsymbol{y(y \in \{0,1\})}$ pro binární klasifikaci (jako je i predikce aktivních míst). KS filtruje funkce objektu, dokud není dosaženo konečného listu (přidělení do třídy). Každý vnitřní bod testuje pouze jeden objekt a jednu jeho funkci $\boldsymbol{v_i}$. Ta se v testu porovnává s prahem $\boldsymbol{\lambda_i}$. Objekt s větší hodnotou dané funkce (resp. menší) bude filtrován doleva (resp. doprava) na „nižší“ uzel. KS se konstruuje na základě trénování vzorků. Začíná se od základního uzlu se všemi vybranými vzorky $\boldsymbol{\left\{\left(v^j, y^j\right), j = 1...L\right\}}$ a roste rekurentně tak, aby se v každém dalším uzlu rozdělil na dva další s maximální třídou homogenity podle rozhodovacího pravidla. Rozhodovací pravidlo hledá vždy nejlepší vlastnosti a nejlepší práh $\boldsymbol{\lambda_i}$ tak, aby maximalizoval informační zisk. Pomocí popsaných pravidel se KS pěstuje až do maximální homogenity. Ukázka KS je na obr. 6. [5]

RF je souborový klasifikátor kombinující N klasifikačních stromů. Každý strom je konstruován užitím náhodné podmnožiny vzorků z původní skupiny. Klasifikace vstupu je získána sečtením (seskupením) hlasů (výstupů) jednotlivých KS. Spojením těchto dvou zdrojů náhodnosti (náhodný výběr jedinců, náhodný výběr funkcí pro stanovení kritérií) se výkon klasifikace pomocí RF výrazně zvyšuje v porovnání s jediným KS.[5]



Obr. 6: Příklad klasifikačního stromu. Zde jsou dvě třídy dat v souboru, v dvourozměrné funkci prostoru (vlevo). Vzorky každé třídy reprezentovány jako čtverce a kruhy (vlevo). KS je postaven vpravo. Pevné linky vlevo ukazují rozdělení prostoru do homogenních regionů. [5]

Frekvenčně získané deskriptory (charakteristiky) budou užity místo nebo spolu s 3D strukturními deskriptory, jako vstup do RF. Z frekvenčních vlastností se použily výše zmíněné 3 největší změny vrcholů spekter, 8 energetických změn na dílčích částech sekvence a globální změna energie. Tyto parametry se použily jak s hodnotami EIIP tak i s IC. Z těchto vlastností se vypočítá 24 různých funkcí. Vybere se ta, která nejlépe rozliší aktivní místa. Při využití RF se ukázaly jako nejlepší 3: největší změny vrcholů spektra (EIIP), energetické změny v 8 dílčích pásmech (EIIP) a změna globální energie (IC). Tyto funkce vytvoří pěti dimensionální prostor zvaný *The sequence-based frequency-derived features in the sequel.* [5]

Pro srovnání jsou zavedeny i 3D strukturní vlastnosti. Jako přístupná povrchová plocha (*Accessible Surface Area*), párový potenciál a výpočet volné vazebné energie. Výpočet volné vazebné energie byl proveden pomocí Robetta serveru. První dvě zmíněné hodnoty (přístupná rozpustná plocha, párový potenciál) byly vypočítány pomocí server HitPoint [9]. [5]

Metoda byla testována na dvoutřídovém datasetu obsahujícím 221 residuí (76 aktivních míst a 145 neaktivních míst). Jako aktivní místa jsou označeny residua s $\Delta\Delta G$ vyšší než 2,0 kcal/mol a jako neaktivní místa ty $\Delta\Delta G$ menším než 0,4 kcal/mol.[5]

Po nastavení kompromisu mezi detekcí aktivních míst a přesností (využívá se F1 měření a Matthewsův korelační koeficient) se získají zajímavé výsledky. Sekvenčně frekvenční přístup *-1DFrekv* - (zmíněné vypočítané frekvenční vlastnosti, po klasifikaci RF) je lepší něž identifikace pomocí 3D strukturních vlastností. A to konkrétně 1DFrekv

detekoval 59% aktivních míst oproti 54% u 3D strukturních, a i více přesněji 75% proti 67%. V kombinaci s 3D strukturních zvedá 1Dfrekv její úspěšnost predikce na 82% s přesností 80%.[5]

3.3.2. Metoda využívající S-Transformace

Tato metoda (publikovaná v [3]) využívá také převodu sekvence na digitální signál. Ale využívá jen hodnoty EIIP pro převod. Dále využívá *Resonant Recognition Model* (dále jen RRM), česky *Rezonanční Rozpoznávací Model.*. [3]

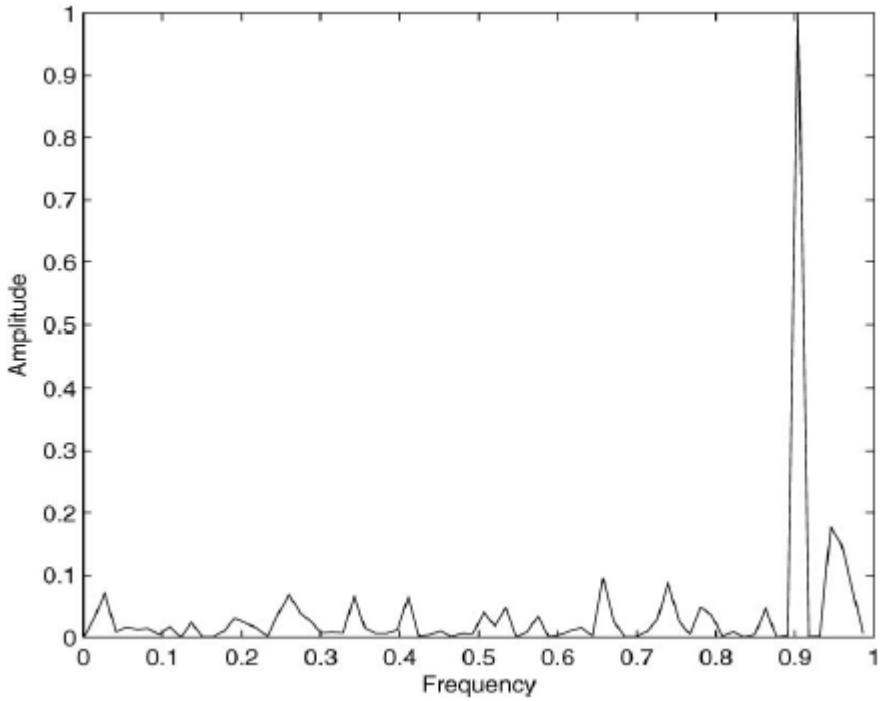
RRM využívá faktu, že biologická funkce je primárně určena modelem protein-cíl interakce. V matematicko fyzikálním přiblížení lze interpretovat proteinovou sekvenční informaci pomocí nástrojů pro zpracování signálů. Je zde významná korelace mezi spektry číslicové reprezentace sekvence aminokyselin s jejich biologickou aktivitou. V [10] je prokázáno, že Fourierova spektrální transformace převedeného proteinu na signál (pomocí EIIP hodnot) má silnou relevanci na sekvenci aminokyselin. Ukázalo se, že všechny proteiny patřící do funkční rodiny sdílí společný spektrální komponent, který charakterizuje jednotlivé funkce skupiny. Tedy je zde charakteristická frekvence skupiny. Protein i cíl oba sdílejí stejnou charakteristickou frekvenci, ale oba v opačné fázi. RRM má dvě fáze. Nejdříve musíme sekvenci aminokyselin převést na číslicový signál (využití hodnot EIIP). Poté na tomto signálu provést Diskrétní Fourierovou Transformaci (dále jen DFT), aby se mohlo provést hodnocení shody spekter. Spektrum funkční rodiny proteinů sdílí podobné frekvence (výrazné vrcholy v amplitudovém spektru).

Pro analyzovanou skupinu proteinů se počítá společné amplitudové spektrum. Jedná se o společné spektrum všech analyzovaných proteinů. Pokud se v tomto spektru vyskytne výrazný vrchol (výrazná společná frekvence) , znamená to že skupina analyzovaných proteinů má společnou biologickou funkci. Výpočet společného spektra:

$$S(\omega) = |X_1(\omega)||X_2(\omega)|...|X_K(\omega)| \quad . \tag{6}$$

Kde $\boldsymbol{X_1, X_2, \ldots X_K}$ jsou DFT korespondujících proteinů (resp. jejich signálů), a S(ω) je společné spektrum. Pokud mají proteiny pouze jedenu společnou funkci, bude pouze jeden vrchol ve společném spektru. Když je více společných funkcí, je i více vrcholů ve společném spektru. (viz. obr.7) [3]

Určením charakteristické frekvence RRM u skupiny proteinů přes společné spektrum lze identifikovat individuální aminokyselinu, které k tomu nejvíce přispívají. Zde by se mohla nacházet aktivní místa zodpovědná za vazbu mezi těmito proteiny. Postup detekce aktivních míst je následující. Mění se postupně amplituda Fourierových koeficientů pro charakteristickou frekvenci. Tím se určí AK, která je na tyto změny nejvíce náchylná, a ta patří k charakteristické frekvenci. Problémem je, že změna jednoho koeficientu působí na všechny AK. To je důvod nutnosti společné časově frekvenční analýzy pro analyzování změny charakteristické frekvence. [3]



Obr. 7: Společné amplitudové spektrum Fibroblástového Růstového Faktoru, vrchol koresponduje s charakteristickou frekvencí relevantní pro určitou biologickou funkci,[3]

Jsou různé možnosti společné časově frekvenční analýzy. Například Fourierova transformace s krátkým časem STFT (STFT – *Short Time Fourier Transformation*). Ta má ale špatné frekvenční rozlišení na nízkých frekvencích a nízké časové rozlišení na vysokých frekvencích. Oba problémy jsou způsobeny pevnou šířkou okna. Další možností je kontinuální vlnková transformace CWT (CWT – *Continuous Wavelet Tranformation*). Ta je lepším provedením společné časově frekvenční analýzy, ale produkuje nevhodné časové úseky pro vizuální analýzu, způsobuje artefakty spektra a absolutní fázová informace nemůže být odvozena. Nejlepší volbou se ukázala S-Transfomace, ta kombinuje výhody STFT a CWT. Poskytuje frekvenčně závislé rozlišení při zachování přímého vztahu s Fourierovým spektrem. S-Transformace signálu x(t):

$$S(\tau, f) = \int_{-\infty}^{\infty} x(t)\omega(\tau - t, f)e^{-j2\pi ft}dt\,, \tag{7}$$

kde $\boldsymbol{\omega(\tau\text{-}t,f)}$ (8) je funkcí okna (škálovatelný gaussian), $\boldsymbol{t}$ představuje čas, $\boldsymbol{f}$ představuje frekvenci spektrální lokalizace a $\boldsymbol{\tau}$ dobu spektrální lokalizace. Tím řeší problém STFT s pevnou délkou okna.

Funkce okna je:

$$\omega(t,\sigma) = \frac{1}{\sigma(f)\sqrt{2\pi}} e^{-\frac{t^2}{2\sigma^2(f)}}, \tag{8}$$

kde

$$\sigma(f) = \frac{1}{|f|}, \tag{9}$$

tedy kombinací (8) a (9) dostáváme

$$\omega(t) = \frac{|f|}{\sqrt{2\pi}} e^{-\frac{t^2 f^2}{2}}. \tag{10}$$

A kombinací (7) a (10) dostáváme:

$$S(\tau, f) = \int_{-\infty}^{\infty} x(t) \left\{ \frac{|f|}{\sqrt{2\pi}} e^{-\frac{(\tau-t)^2 f^2}{2}} e^{-j2\pi f} \right\} dt. \tag{11}$$

Výsledná funkce okna (10) by se dala rozdělit na dvě části. Jedna část je pomalu se měnící obálka, která lokalizuje čas. Druhá část je oscilační jádro, které volí frekvence. Lokalizuje fázi jako amplitudové spektrum. Tak nám zůstane absolutní fázová informace, což není k dispozici u CWT. [3]

U společné časově frekvenční analýzy je problém s filtrováním. U standardní Fourierovy transformace je filtrování pomocí fixní pásmové propusti po celou dobu. Pro společné časově frekvenční analýzy je ale zapotřebí filtry s časově proměnou pásmovou propustí. Nejdříve se ale musí provést odhad částí co jsou šum, a které jsou užitečný signál. První odhad lze provést pomocí skupiny pásmových zádrží, které odstraní uvažovaný šum.
Okno S-Transformace okno splňuje podmínku:

$$\int_{-\infty}^{\infty} \omega(t, f) dt = 1, \tag{12}$$

proto průměrování ***S(τ,f)*** přes všechny hodnoty ***t*** dává ***X(f)***, což je FT ***x(t)***.

$$\int_{-\infty}^{\infty} S(\tau, f) d\tau = X(f). \tag{13}$$

Proto lze získat původní signál inverzní FT z ***X(f)***:

$$x(t) = \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} S(\tau, f) d\tau \right\} e^{j2\pi ft} df. \tag{14}$$

Tento vztah (14) nám poskytuje přímý vztah mezi S-Transformací a FT. Ten lze vhodně využít pro časově frekvenční filtrování. Kdy ***x(t)*** = ***d(t)*** + ***n(t)***, kde ***x(t)*** je celý signál, ***d(t)*** užitečný signál a ***n(t)*** je šum. Díky tomu vzhledem k linearitě lze zapsat
S(τ,f) = ***D(τ,f)*** + ***N(τ,f)***, kde ***D(τ,f)*** a ***N(τ,f)*** jsou S-Transformací užitečného signálu, resp.šumu.

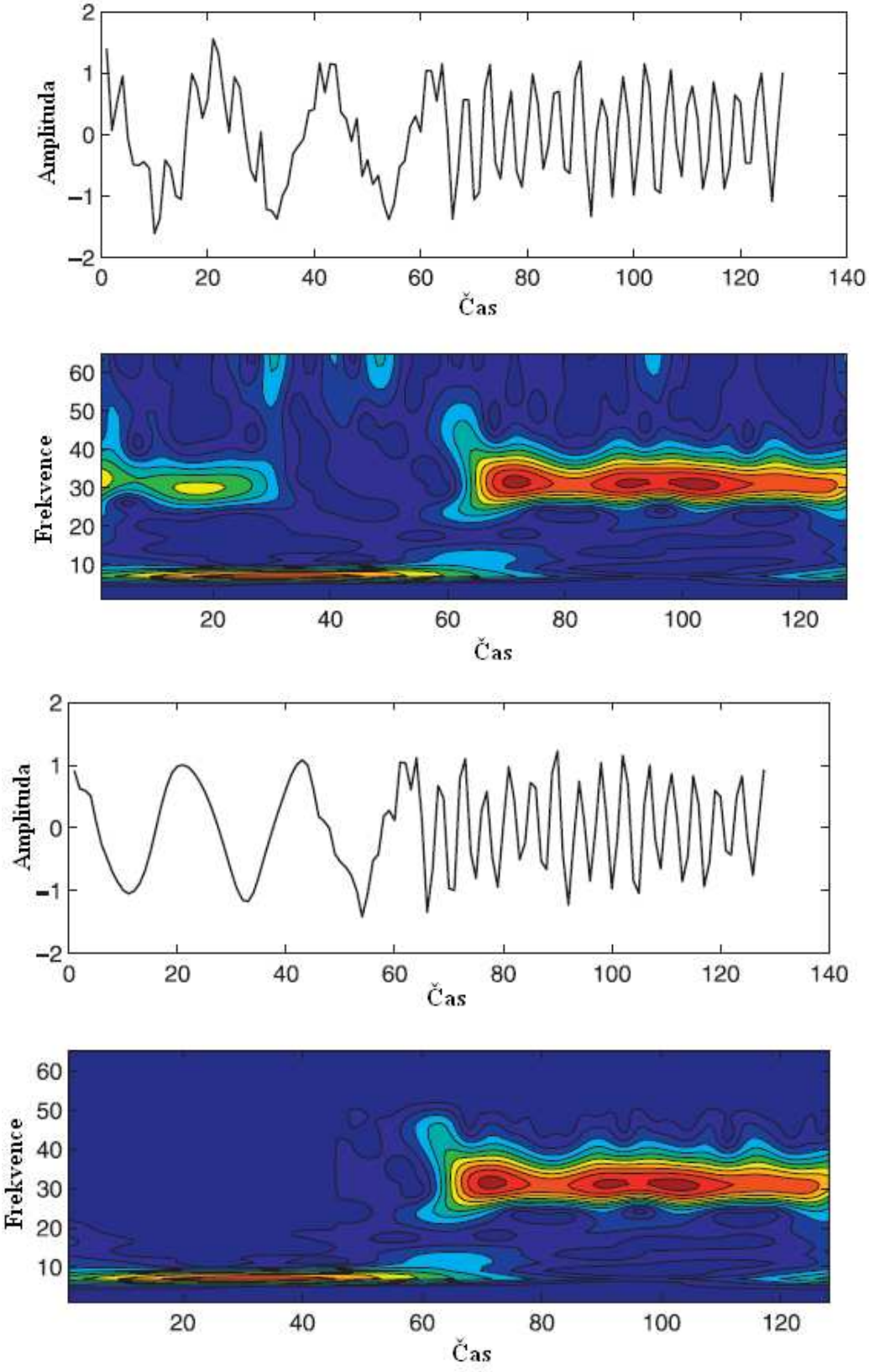
Tím pádem filtrační funkce ***A(τ,f)*** je zvolena tak, aby ***D(τ,f)*** = ***A(τ,f).S(τ,f)***. Užitím inverze dostáváme bezšumný signál:

$$\tilde{x}(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} D(\tau, f)e^{j2\pi ft}d\tau df = \int_{-\infty}^{\infty} \tilde{X}(f)e^{j2\pi ft}df\,, \qquad (15)$$

kde

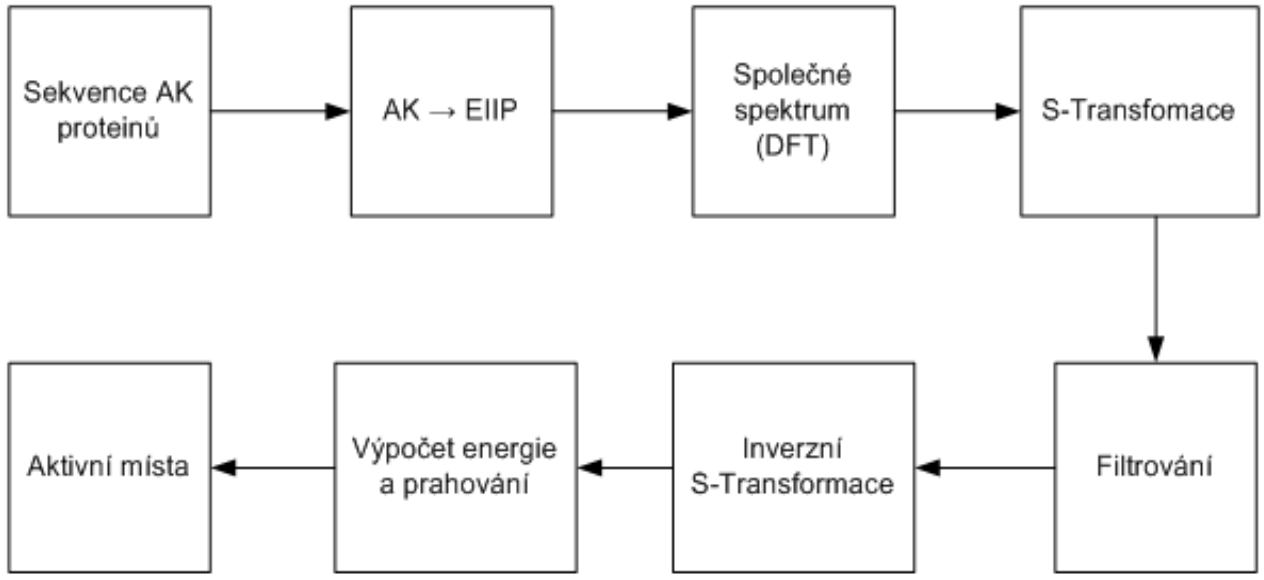
$$\tilde{X}(f) = \int_{-\infty}^{\infty} D(\tau, f)df\,, \qquad (16)$$

proto vynásobení S(τ,f) s filtrační funkcí A(τ,f) dává bezšumný signál. Pro STFT tento postup nelze provést, není pro ni inverzní operace z časově frekvenční roviny. Filtrovaná S Transformace může být velmi užitečná při identifikaci aktivních míst. Přínos filtrace na obr.8. [3]



Obr. 8: Ukázka filtrace s využitím S-Transformace, nahoře je zašuměný signál a jeho S-Transformace a pod ním je signál a jeho S-Transformace po filtraci,[3]

Při využití S-Transformace pro predikci aktivních se postupuje následně. Nejdříve se aminokyseliny převedou na diskrétní signál (pomocí EIIP hodnot). Jelikož EIIP hodnoty jsou od 0 do 0,1263, vzniká stejnosměrná složka (průměrná hodnota signálu). Ta však nemá v kontextu spektrální analýzy význam. Mohlo by navíc dojít ke stejnosměrnému ovlivnění (klamavé špičky ve spektru). Proto se stejnosměrná složka odstraňuje ještě před DFT. Poté je vypočítáno společné spektrum (6). Vrchol ve spektru koresponduje s charakteristickou frekvencí, nelze ale identifikovat jednotlivé aminokyseliny s maximálním přínosem na daný vrchol. Z tohoto důvodu se provede společná časově frekvenční analýza. Spektrum se počítá pro rozpoznání rozdělení energie charakteristické frekvence po celé délce sekvence. Zkoušely se i jiné metody jako STFT a CWT, ale S-Transformace má lepší rozlišení. Tím pádem lépe odhalí energeticky významné oblasti v proteinu pro danou frekvenci. Za účelem snížení šumu a zvýraznění energie charakteristické frekvence společného spektra se násobí spektrum S-Transformace se společným spektrem na každém vzorku (vždy spektrum jednoho signálu se společným spektrem). Po nalezení vrcholů ve spektru (lokální maxima) jsou porovnávány, resp. jejich poměr s průměrnou energií. Pokud je tento poměr větší než stanovený práh, je tam určeno aktivní místo (v této metodě je výchozí stanoven práh roven jedné). Schéma metody je vidět na obr.9. [3]



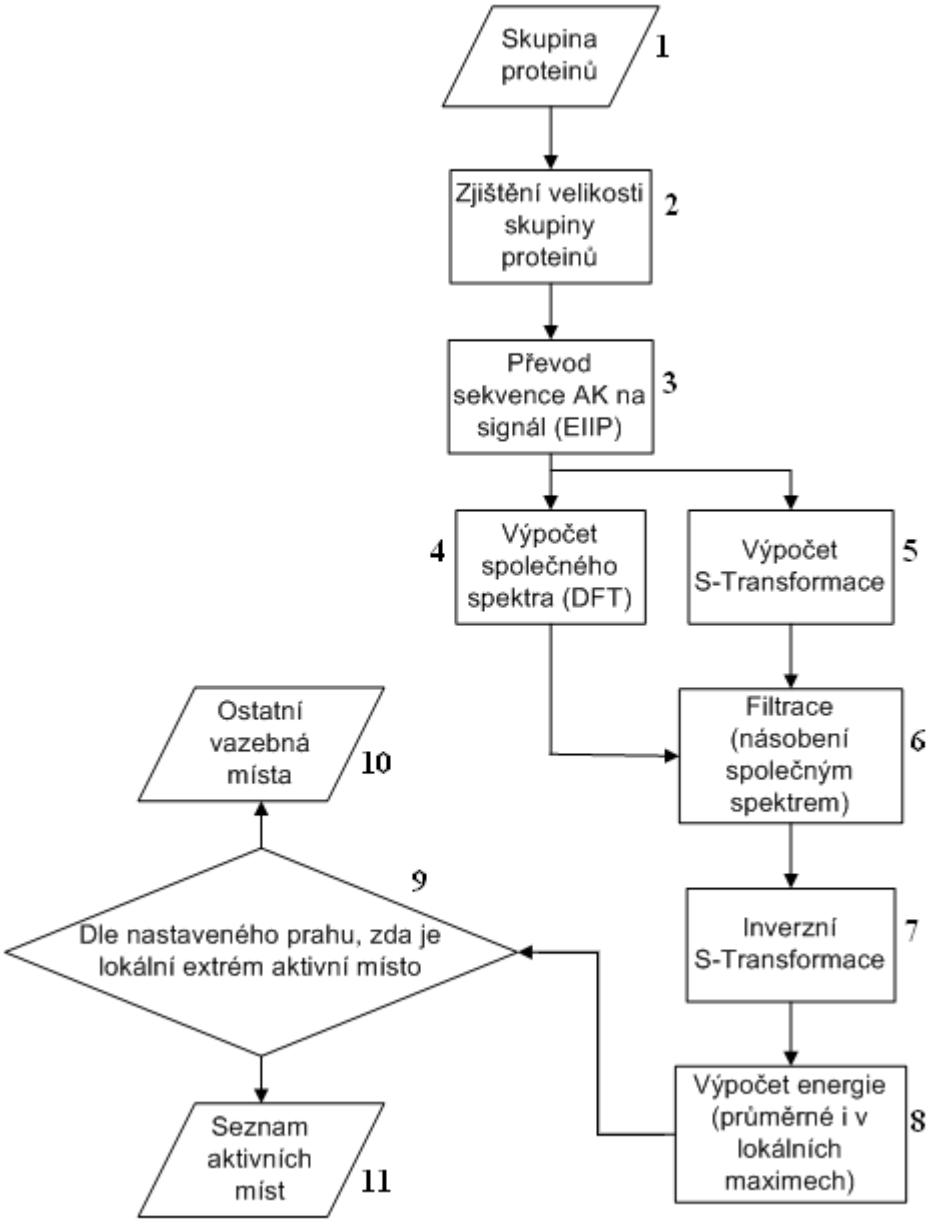
Obr. 9: Schéma metody s S-Transformací, [3]

Tato metoda dosahuje úspěšnosti sensitivity cca. 79% což je více oproti metodám založeným na digitálním filtrování, ty jen 67%. Tato metoda je úspěšnější než jiné metody, např. KFC server nebo HotPoint. Má také ale své nevýhody: dává falešně pozitivní hodnoty, nízké frekvenční rozlišení na vysokých frekvencích, malé časové rozlišení na nízkých frekvencích (to je způsobeno gaussianem). Pokud by se podařilo zlepšit masku (filtr) v časově frekvenční rovině, mohlo by to pomoci snížit počet falešně pozitivních predikcí. [3]

4. Návrh pseudokódu

4.1. Vývojový diagram

Pro návrh pseudokódu jsem si vybral metodu popsanou v bodě 3.3.2. Tedy metodu využívající S-Transformaci. Pro lehčí orientaci v návrhu pseudokódu jsem nejdříve vytvořil vývojový diagram programu (obr. 10).



Obr. 10: Vývojový diagram metody s S-Transformací

Kdy jako vstup {1} je skupina zarovnaných proteinů stejné délky. Pokud by tak nebylo muselo by se provést jejich případné předzpracování s cílem proteiny zarovnat a znormovat jejich délku na stejnou hodnotu. V dalším bloku {2} se zjistí velikost skupiny proteinů

(celkový počet proteinů a jejich délka). V bloku {3} se provede převod jednotlivých sekvencí proteinů reprezentovaných posloupnostmi aminokyselin (případně lze program upravit tak aby nejdříve z posloupnosti nukleotidů vytvořil posloupnost aminokyselin) převod na skupinu diskrétních signálů. Pro převod se využijí hodnoty EIIP. Blok {4} provede pro všechny signály (resp. sekvence proteinů) DFT a poté jejich vzájemné vynásobení jejich amplitudových spekter. Tím se získá společné spektrum. V bloku {5} se ze skupiny signálů (stejná jako byla užita v bloku {4}) spočítá jejich S-Transformace podle vzorce (11). Blok {6} je asi nejdůležitější. Zde se provede za prvé filtrace a zadruhé normování. Filtrace spočívá v použití časově frekvenčních filtrů (bude obtížný jejich návrh, z hlediska toho co je a není užitečný signál). Normování znamená násobení spekter S-Transformací jednotlivých signálů společným spektrem (vzorek po vzorku). To potlačí šum a zvýrazní společnou frekvenci. V {7} se provede inverzní S-Transformace dle (15). V bloku {8} se spočítá energie lokálních maxim a průměrná energie signálu. Tyto hodnoty se v bloku {9} použijí pro jejich vzájemný poměr (energie lokálního maxima vůči energii průměrné). A tento poměr se porovná s prahem, dle tohoto porovnání se určí zda místo putuje do {10} (menší než práh) nebo do bloku {11} (větší než práh).

4.2. Pseudokód

Nejdříve popíší samotný pseudokód pod ním se nalézá vysvětlení jednotlivých funkcí a tabulka proměnných.

Vyhledávaní aktivních míst (proteiny,EIIP,filtr,práh)

```
1)   [m,n]← Zjisti velikost(proteiny)
2)   For i ← 1 až m
3)       sekvence(i) ← protein(i)
4)   signály ← Převod sekvence(sekvence,EIIP)
5)   For i ← 1 až m
6)       spektra(i) ←DFT(signály(i))
7)   S_spektrum← signály(1)
8)   For i ← 2 až m
9)       S_spektrum← S_spektrum*spektra(i)
10)  ST_spektra← S-Transformace(signály)
11)  For i ← 1 až m
12)      Filt_ST_spektra(i) ← filtr*ST_spektra(i)
13)      For j ← 1 až n
14)          Filt_ST_spektra(i,j)← Filt_ST_spektra(i,j) )*S_spektrum(j)
15)  Filt_signály← Inv_S-Transformace(Filt_ST_spektra)
16)  Pozice_maxim← Pozice(Filt_signály)
17)  if délka(Pozice_maxim) ≥ 1
```

18) Energie_maxim←Výpočet energie 1(Filt_signály,Pozice_maximum)
19) Energie_průměr ←Výpočet energie 2(Filt_signály
20) [Aktivní_místa,Ostatní_místa]←
Rozhodnutí(Pozice_maxim,Energie_maxim,Energie_průměr,práh)
21) Konec

Toto je návrh pseudokódu metody popsané v 3.3.2. Nyní vysvětlím funkce jednotlivých funkcí. Význam jednotlivých proměnných je uveden v tabulce 7.

Zjisti velikost(proteiny) zjistí velikost a počet proteinů. Jak bylo zmíněno výše do programu by měla vstupovat skupina zarovnaných stejně dlouhých proteinů. Při zjištění že tomu tak není ohlásí funkce chybu. Tím se ukončí běh celého programu a uživatel bude upozorněn na nutnost předzpracování proteinů.

Převod sekvence(sekvence,EIIP) tato funkce každé jednotlivé aminokyselině v každém proteinu přiřadí číslo z hodnot EIIP. Tím se vytvoří skupina diskrétních signálů, které se budou dále zpracovávat. Tuto funkci lze upravit i tak aby nejdříve z posloupnosti jednotlivých nukleotidů vytvořila posloupnost aminokyselin a tu poté převedla na diskrétní signál.

DFT(signály) tato funkce nejdříve ze všech signálů odstraní nežádoucí stejnosměrnou složku (průměrnou hodnotu). A poté provede pro všechny signály diskrétní Fourierovu transformaci. Jejím výstupem je skupina amplitudových spekter odpovídajících vstupním signálům (bez stejnosměrné složky).

S-Transformace(signály) tato funkce provede S-Transformaci signálů. Tu provede podle vzorce popsaného v (11). Výstupem je S-Transformace vstupních signálů.

Inv_S-Transfomace(Filt_ST_spektra) tato funkce provede inverzní S-Transformaci na základě vzorců (15) a (16).

Pozice(Filt_signály) najde lokální maxima signálů (peaky). K tomu lze využít například vzorce (1).

Délka(Pozice_maxim) je funkce podobná funkci *Zjisti velikost* ale zde se zjišťuje pouze „délkový“ rozměr. Tedy počet hodnot v proměnné.

Výpočet energie 1(Filt_signály,Pozice_maximum) tato funkce vypočítá energii lokálních maxim(vrcholů) signálů. Energie se vypočítá podle vzorce:

$$\boldsymbol{E(n)} = |\boldsymbol{y(n)}|^2, \tag{17}$$

kde $\boldsymbol{E(n)}$ je energie vzorku a $\boldsymbol{y(n)}$ je výstupní filtrovaný signál.

Výpočet energie 2(Filt_signály) zde se vypočítá průměrná energie každého signálu.

Rozhodnutí(Pozice_maxim,Energie_maxim,Energie_průměr,práh) tato funkce vypočítá poměr mezi energii lokálního maxima (peaku) vůči průměrné energii. A tuto hodnotu porovná s hodnotou prahu. Ta je prvotně nastavena na 1. Pokud je tato hodnota větší než práh, potom je místo označeno za aktivní místo. Pokud ne, tak místo není označeno za aktivní místo.

Tab. 7: Význam proměnných z pseudokódu

Jméno proměnné	Význam prměnné	Typ proměnné
proteiny	skupina zarovnaných stejně dlouhých proteinů	Matice buněk
m	celkový počet proteinů	Celé číslo
n	délka proteinů (počet aminokyselin)	Celé číslo
sekvence	sekvence AK jednotlivých proteinů	Matice rozměru m x n
i,j	pomocné proměnné pro for cykly	Celá čísla
EIIP	hodnoty EIIP pro jednotlivé aminokyseliny	Matice
signály	skupina signálů vycházejících z proteinů	Matice rozměru m x n
spektra	Ampltiudová spektra signálů	Matice rozměru m x n
S_spektrum	společné spektrum	Vektor délky n
ST_spektra	S-Transformace signálů	Matice rozměru m x n
filtr	časově frekvenční filtr	Vektor
Filt_ST_spektra	filtrované ST_spektra	Matice rozměru m x n
Filt_signály	výstupní filtrovaný signál	Matice rozměru m x n
Pozice_maxim	pozice lokálních maxim (peaků)	Vektor
Energie_maximum	energie lokálních maxim (peaků)	Matice
Energie_průměr	průměrné energie každého signálu	Vektor délky n
práh	práh pro určení aktivního místa	Realné číslo
Aktivní_místa	seznam pozic aktivních míst	Vektor
Ostatní_místa	seznam pozic co nejsou aktivní místa	Vektor

5. Softwarové řešení

Pro technické řešení bylo určeno programové prostření Matlab. V tomto prostředí jsem zvolil přístup pomocí funkcí (prostřednictvím tzv. m-file souborů, tedy souborů s koncovkou ***.m). Kde jsem nejdříve vytvořil hlavní soubor *METODA.m* ze kterého se postupně volají další funkce (soubory) aby provedly posloupnost úkolů vedoucích k požadovanému výsledku. Tedy určení aktivních míst za využití společného spektra a S-Transformace.

5.1. Popis hlavního souboru METODA.m

Jak bylo zmíněno výše hlavní soubor (funkce) nese název *METODA.m*, v této funkci proběhnou všechny operace (jako volání vedlejších funkcí). Proto lze proměnné v této funkci označit jako globální proměnné (jsou uvedeny v tabulce 8).

Tab. 8: Význam proměnných

Jméno proměnné	**Význam proměnné**	**Typ proměnné**
sekvence	do této proměnné se postupně načtou jednotlivé sekvence ve formátu fasta	Matice buněk (typ cell)
n	rozměr matice buněk, počet zpracovávaných sekvencí	skalární proměnná
delka	vektor délek jednotlivých sekvencí	vektor
signaly	signály získané převodem sekvencí na signály	matice buňek
protein_c	pro ktetrou sekvenci se má provést výpočet S-Transformace	skalární proměnná
spektrum	společné spektrum signálů (sekvencí)	vektor
yy	signál získaný po S-Transformaci, pronásobení se společným spektrem a inverzní S-Transformaci	vektor
en	vektor energií jednotlivých prvků *yy*	vektor
prum_en	průměrná enregie signálu *yy*	skalární proměnná
pozice	vektor pozic na kterých bych se mohly vyskytovat aktivní místa	vektor
overene_pozice	vektor ověřených pozic označených jako aktivní místa	vektor

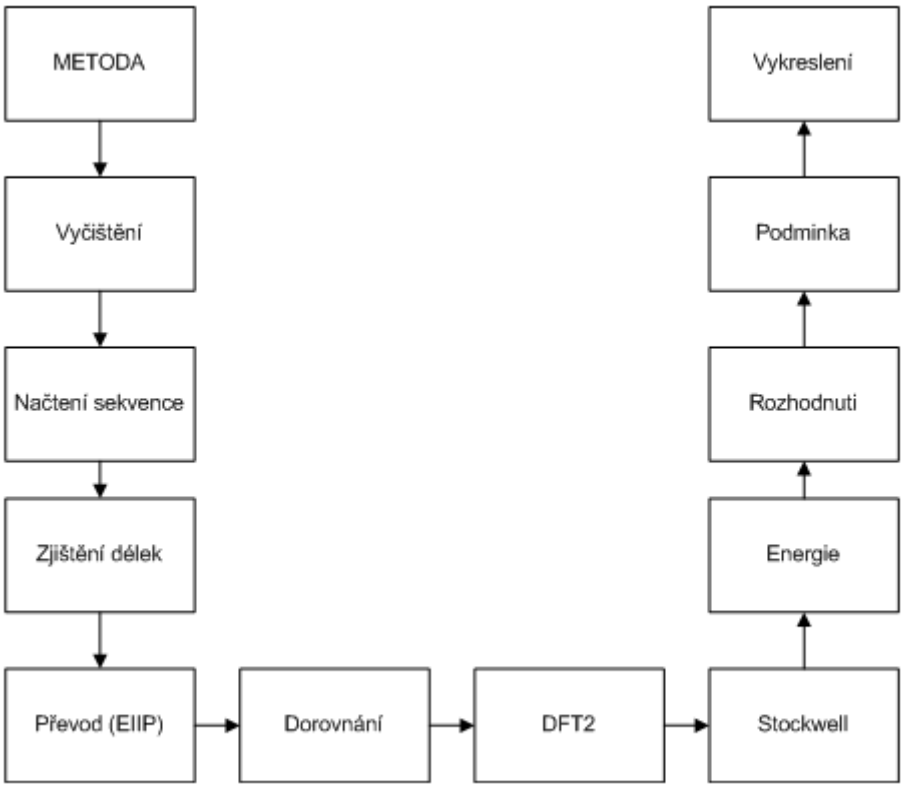
V této tabulce ještě chybí proměnné *m* (není v programu dále využita) a proměnné *w,ww,h1,h2,h4* ty jsou pouze využity pro vykreslení výslekdů.

Funkce *METODA.m* nejdříve pomocí trojice příkazů vše „vynuluje“. Tedy vyčistí příkazový řádek, vymaže všechny dosud uložené proměnné a zavře všechna aktuálně otevřená okna (programu Matlab). Poté příslušným příkazem

načte sekvence a uloží je do proměnné *sekvence*. Z této proměnné zjistí daným příkazem počet sekvencí (proměnná *n*). Pomocí for cyklu zjistí délku jednotlivých sekvencí a uloží je jako vektor do proměnné *delka*. Převod sekvencí na signál za využití hodnot EIIP je zajištěn příkazem zavoláním funkce pro převod sekvencí na signály. Ale tyto signály nejsou většinou stejně dlouhé (různá délka původních sekvencí) a pro další zpracování je nutné aby měli stejný rozměr. Já jsem se rozhodl je zarovnat na maximální délku, která se v souboru signálů vyskytne. Jako data na „dorovnání“ jsem použil průměrné hodnoty každého jednotlivého signálu, čímž tedy dostanu soubor stejně dlouhých signálů vhodných k dalšímu zpracování. Předtím ještě pomocí proměnné *protein_c* určím pro který protein ze souboru se budou určovat aktivní místa. Primárně by to měl být první protein ze souboru. Zpracování signálu začíná voláním funkce *DFT2* (soubor *DFT2.m*) v této funkci se spočítá společné spektrum dle vzorce (6). Přesné fungování bude vysvětleno dále. Když je vypočítáno společné spektrum program začne vypočítávat S-Transformaci její násobení se společným spektrem a zpětnou S-Transformaci pomocí funkce *stockwell* (soubor *stockwell.m*). Z výsledného signálu z této funkce se spočítá jeho energie dle vzorce (17) a jeho průměrná energie podle vzorce

$$E_P = \frac{1}{N}\sum_{n=1}^{N}\left|x(n)\right|^2, \tag{18}$$

kde $\boldsymbol{E_P}$ je průměrná energie, $\boldsymbol{N}$ je počet vzorků signálu a $\boldsymbol{x(n)}$ je vzorek signálu. Tato funkce se jmenuje *energie* (soubor *energie.m*). Po té se musím rozhodnout z poměru energie vzorku a jeho poměrné energie zda by se zde mohlo nacházet aktivní místo. Zde jsem použil předpoklad, že aktivní místo se nenachází pouze na jedné aminokyselině. Ale na minimálně dvou (určených jako potenciálních) poblíž sebe. Tedy alespoň dvě ze tří určených pozic musí být poblíž sebe. To znamená (pro názornost si určíme, že 1 znamená určenou pozici a 0 nikoli) tyto možnosti 110, 011 a 101. A samozřejmě i větší části sekvencí pokud budou soustavně splňovat tyto podmínky. Tento problém řeší funkce *podmínka* (soubor *podmínka.m*). Z této funkce dostaneme ověřené pozice, které již jen v závěru pomocí sekvence příkazů zobrazím. Tento sled příkazů vykreslí vektor pozic a jejich ověření. Poté ještě zavedu pomocný nulový vektor *ww* o délce původní zkoumané sekvence. Do kterého pomocí for cyklu a jednoduché podmínky přiřadím na ověřené pozice hodnoty původního signálu. Tím pádem po vykreslení do společného okna se ukáže kde se na původním signálu nacházejí potenciální aktivní místo. Toto vykreslení je provedeno dalším sledem příkazů. Pro názornost ještě celý algoritmus funkce *METODA* zobrazen na obr. 11.



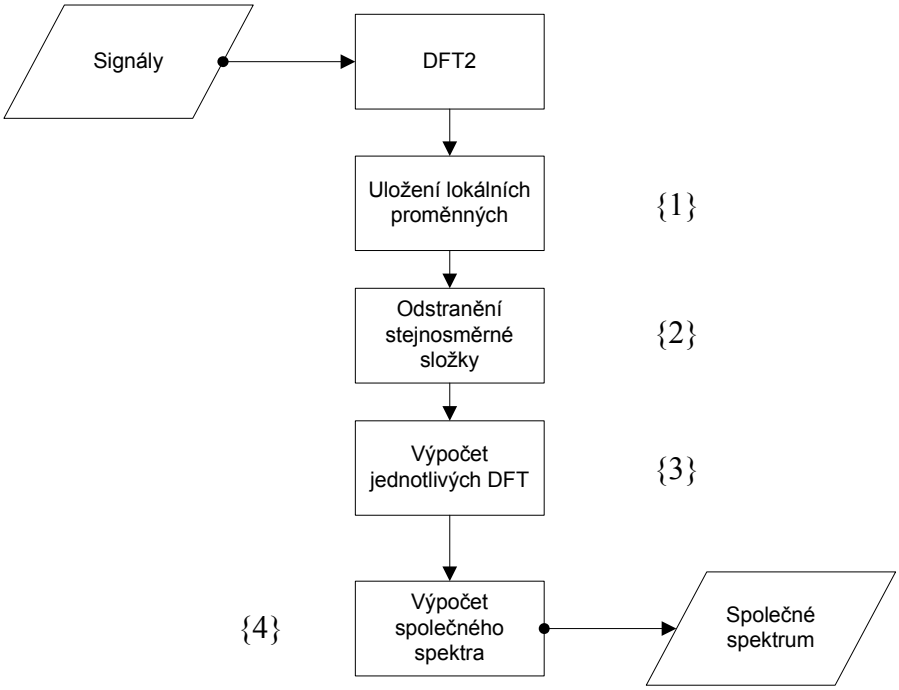
Obr. 11: Algoritmus funkce *METODA*

5.2. Popis vedlejšího souboru DFT2.m

Jak bylo popsáno výše do této funkce vstupuje skupina stejně dlouhých signálů. Ze kterých se bude počítat Diskrétní Fourierova Transformace (DFT) a následně společné spektrum. DFT se bude počítat podle vzorce

$$X[k] = \sum_{n=1}^{N} x(n) e^{-j2\pi kn/N} , \tag{19}$$

kde ***X[k]*** je spektrální koeficient, ***N*** je délka okna (část sekvence, pro kterou se počítá DFT – v mém programu je N rovno délce sekvence), ***k*** je koeficient spektra z intervalu 1 až N/2 (v mém programu ale 1 až N, omezené 1 až N/2 se použije až pro zobrazení), ***j*** je imaginární číslo a ***x(n)*** je ***n***-tá hodnota v sekvenci vymezená oknem N. Společné spektrum se bude počítat podle vzorce (6).



Obr. 12: Algoritmus funkce *DFT2*

Funkce si nejdříve v bloku {1} do proměnných *N,k* (skalární proměnné) uloží maximální délku signálů (zde již délku všech signálů). Samotné signály si uloží do proměnné *x*. A pomocí dvou do sebe vnořených cyklů v bloku {2} odečte od každého signálu jeho stejnosměrnou složku (tedy jeho průměrnou hodnotu). Dále funkce vytvoří pomocnou proměnnou *X* (nulová matice rozměru *n*x*N*). Do ní pomocí tří vnořených for cyklů přiřadí jednotlivé DFT. To se realizuje v bloku {3}. Využívá se proměnná *c* do které se pomocí daného příkazu přiřazuje $e^{-i2\pi kt/N}$, ve kterém proměnná *t* supluje *n* ze vzorce (19), jelikož *n* je již obsazeno jako globální proměnná udávající počet sekvencí. Tyto hodnoty se postupně ukládají do pomocné proměnné *pomoc* (jedná se o nulový vektor délky *N*). Ten se poté pomocí příkazu sečte a uloží na příslušné místo v matici *X* jako příslušný spektrální koeficient.

Společné spektrum se vypočítá pomocí dvou vnořených for cyklů (jsou uvedeny níže) v bloku {4}. Nejdříve se pomocí následujícího příkazu uloží absolutní hodnota prvního řádeku matice *X* tedy amplitudové spektrum prvního signálu (sekvence).Poté již následují zmíněné for cykly. Kde proměnná *s* je následně uložena jako výsledná proměnná této funkce, tedy *spektrum*. Ještě pomocí for cyklu proběhne normování maximální hodnotou. A následuje již jen vykreslení.

Vstupem funkce *DFT2* jsou tedy signály, jejich délky a jejich počet. Výstupem je společné spektrum.

5.3. Popis vedlejšího souboru stockwell.m

V tomto souboru (funkci) se provede S-Transformace, vynásobení se společným spektrem a zpětná S-Transformace.

S-Transformace měla být podle návrhu provedena dle vzorce (11) z [3] resp. jeho přepisu pro diskrétní variantu

$$S[\tau, f] = \sum_{k=1}^{N} x(k) \frac{|f|}{\sqrt{2\pi}} e^{-\frac{(\tau-k)^2}{2}} e^{-j2\pi f}, \tag{20}$$

kde ***S[τ,f]*** je spektrální koeficient, ***τ*** představuje lokalizaci doby, ***f*** představuje lokalizaci frekvence, ***k*** reprezentuje čas (resp. pozici na signálu), ***x(k)***je k-tá hodnota signálu. S tímto vzorcem jsem získal velmi podobné výsledky jako v [3], je znázorněno na obr. č. 13.

Ale nastává problém při pokusu o zpětnou S-Transformaci podle vzorce upět uvedeného v [3] tedy vzorec (14). Opět byla provedena jeho úprava tak aby byl využitelný pro diskrétní data

$$x(k) = \sum_{f=1}^{M} \left\{ \sum_{\tau=1}^{N} S(\tau, f) \right\} e^{j2\pi fk}, \tag{21}$$

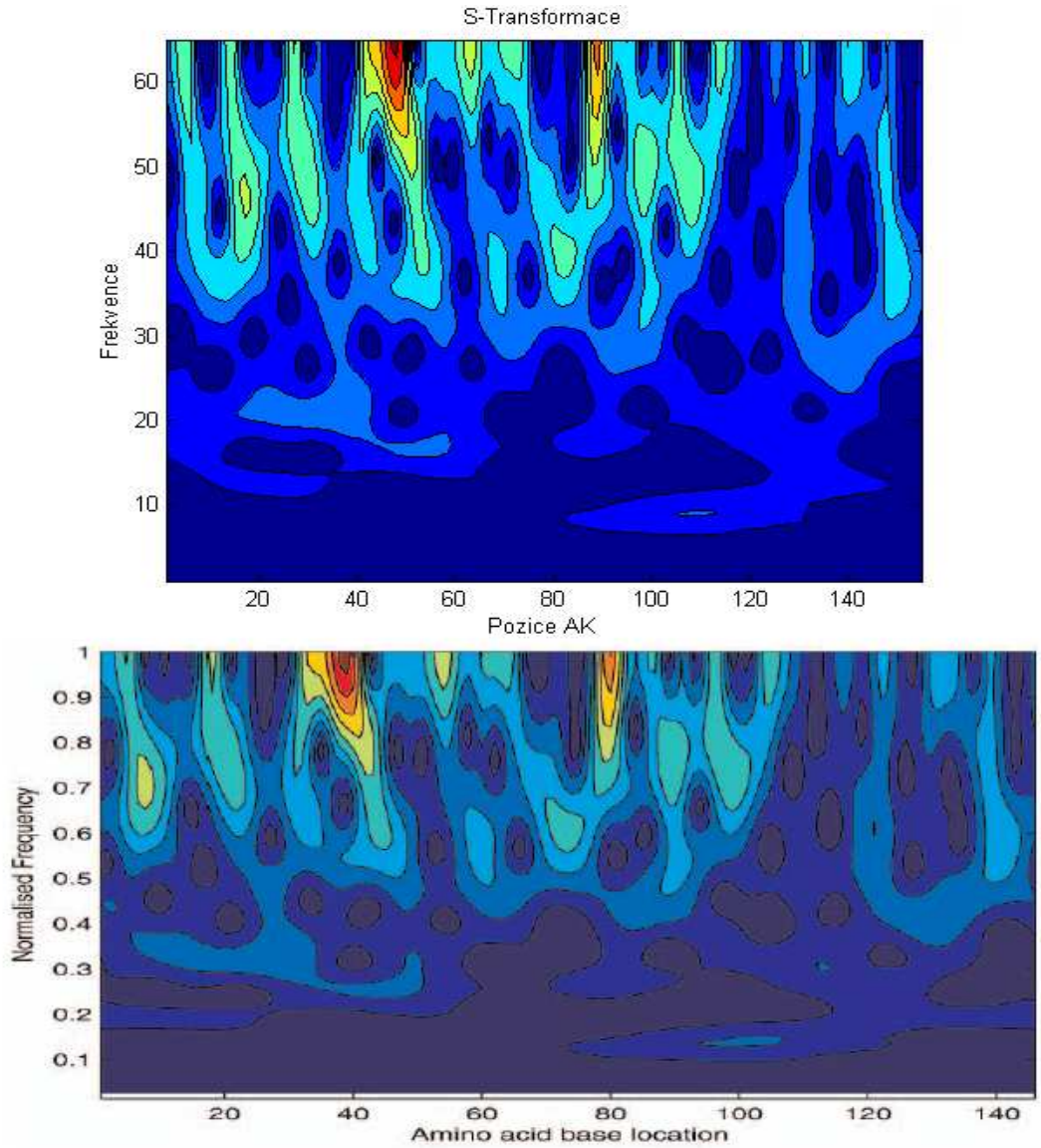
kde ***x(k)*** je původní signál, ***k*** je pozice vzorku na signálu, ***f*** je lokalizace frekvence, ***M*** je počet frekvencí, ***τ*** je lokalizace času, ***N*** je počet frekvencí. Při použití tohoto vzorce jsou ale dostávány nesmyslné výsledky. Zejména proto, že část vzorců (14) a (21) $e^{j2\pi fk}$ dává pouze reálná čísla (tedy nedojde k odstranění imaginární části která vznikne při přímé S-Transformaci). Tím pádem i když se signálem „nic neprovedeme“ tak zpětná S-Transformace dává imaginární hodnoty. Proto jsem se rozhodl použít jiné vzorce, uvedené v [12]. Tedy pro diskrétní S-Transformaci

$$S\left[jT, \frac{n}{NT} \right] = \sum_{m=-N/2}^{N/2-1} H\left[\frac{m+n}{NT} \right] e^{-\frac{2\pi^2 m^2}{n^2}} e^{\frac{i2\pi mj}{N}}, \quad n \neq 0, \tag{22}$$

kde ***H[n/NT]*** je fourierova transformace o N bodech z původního signálu, ***j***, ***m*** a ***n*** jsou z intervalu 0 až N – 1, ***T*** je interval vzorkování (v mé práci je roven 1 – signál se nevzorkuje) a ***i*** je imaginární číslo.A tedy ***S[jT,n/NT]*** je S-Transformace se vzorky pro každý čas a každý frekvenční fourierův vzorek. A pokud se ***n=0*** je S-Transformace definována

$$S[jT, 0] = \frac{1}{N} \sum_{k}^{N-1} h[kT]. \tag{23}$$

Kde ***h[kT]*** je původní signál.



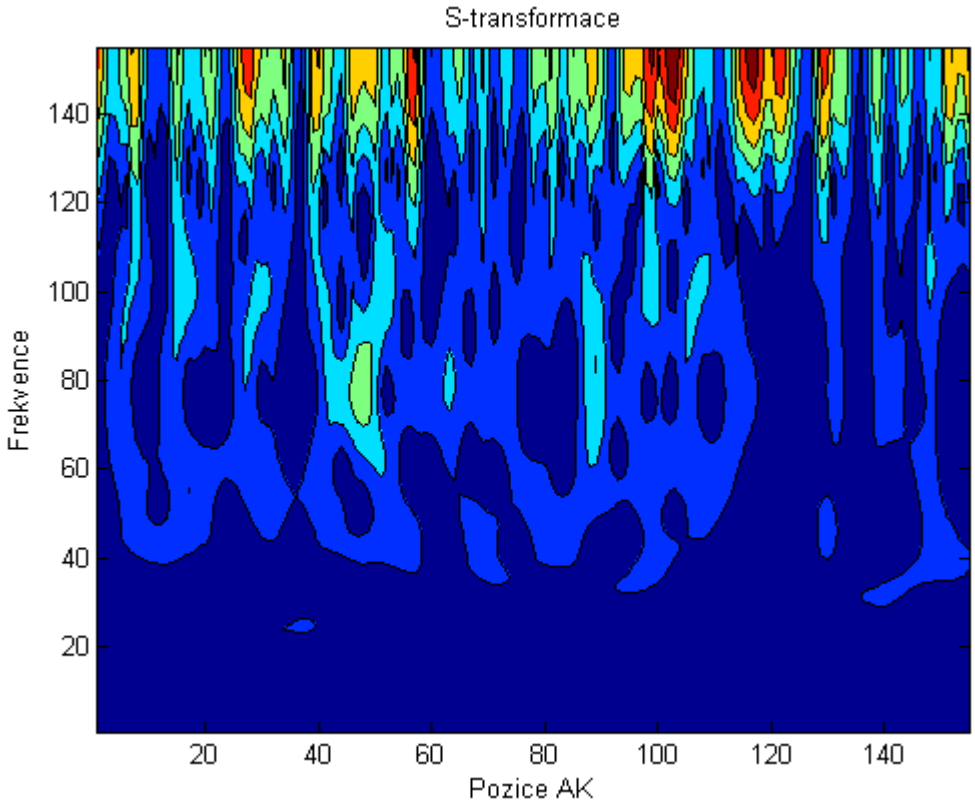
Obr. 13: Porovnání S-Transformací (horní je mnou vytvořená, dolní je převzata z [3] – rozdíl ve frekvencích je dán tím, že jsem je nenormalizoval) pro „*Basic Bovine*“

Diskrétní inverzní S-Transformace se vypočítá dle vzorce

$$h[kT] = \frac{1}{N}\sum_{n=1}^{N}\left\{\sum_{j=1}^{N} S\left[\frac{n}{NT}, jT\right]\right\} e^{\frac{i2\pi nk}{N}}, \qquad (24)$$

kde ***h[kT]*** je obnovený signál, ***N*** je jako ve vzorcích (22) a (23) délka fourierovy transformace a ***S[n/NT,jT]*** je transponovaná matice S-Transformace.

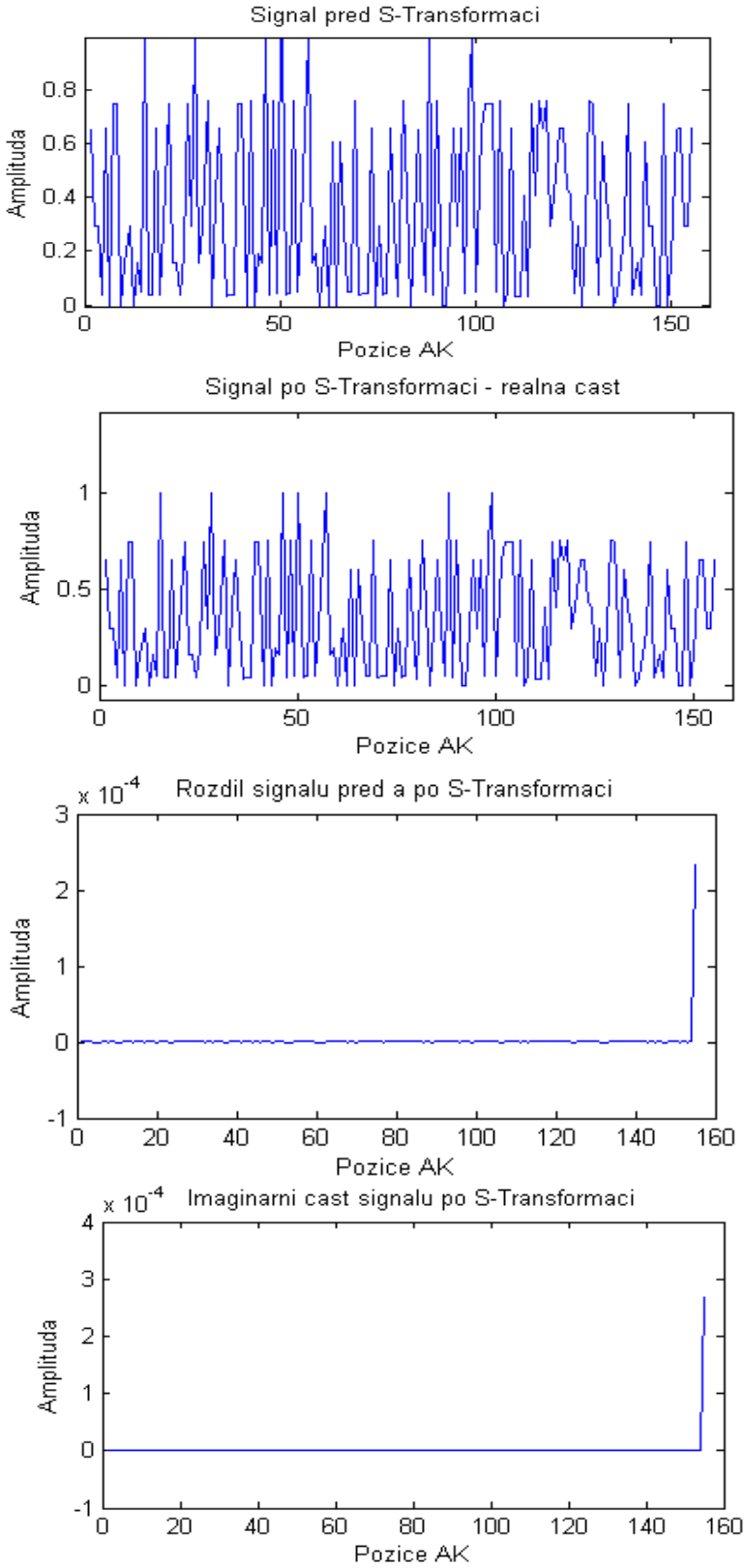
Při použití těchto vzorců dostáváme jiné spektrum S-Transformace (je vidět na obr. 13) , ale pokud se signálem „nic neprovedu“ tak po zpětné S-Transformaci dostávám skoro původní signál. Nepatrný rozdíl je jen na posledním vzorku signálu (to je vidět na obr. 15).



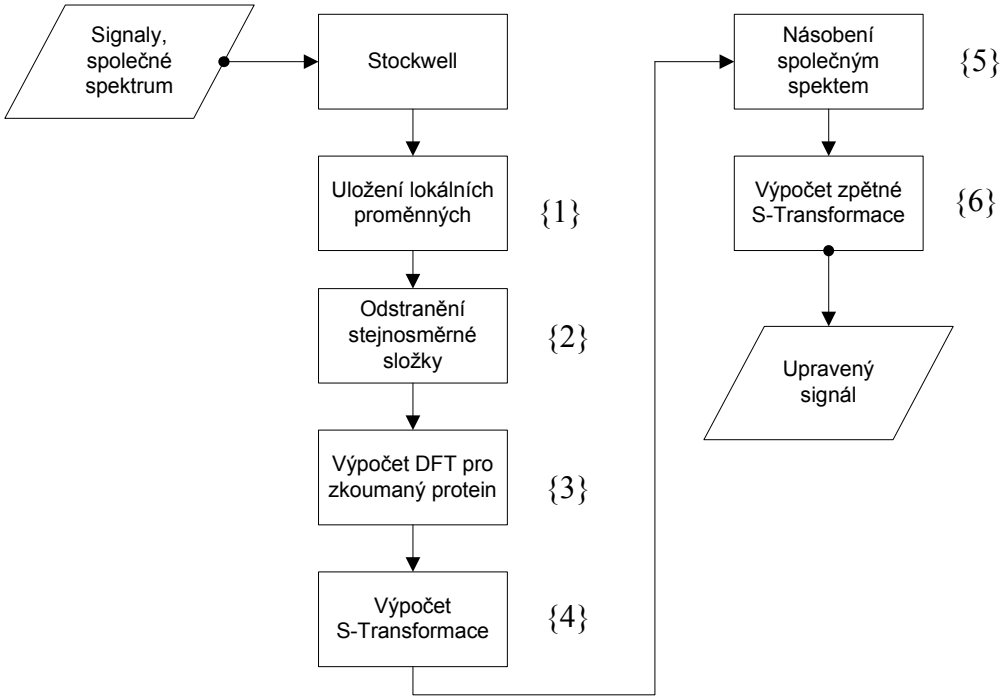
Obr. 14: S-Transformace pro „*Basic Bovine*“, počítána dle vzorce (22)

Jak je vidět na obr. 14 a jak vyplývá ze vzorců (22) a (24) bude mít matice S-Transformace rozměr ***N*** x ***N***. ***N*** je vysvětleno u vzorců (22) a (24).

Když jsem vysvětlil změnu použitých vzorců oproti návrhu konečně se dostávám k samotnému naprogramování souboru. Nejdříve, obdobně jako u souboru *DFT2.m* je uložena délka signálu do proměnné *N* a signály do proměnné *x*, v bloku {1}. Opět je také odečtena stejnosměrná složka (průměrnou hodnotu) ze zpracovávaného signálu, v bloku {2}. Následuje vytvoření pomocných proměnných *Q* a *q*(obě dvě proměnné budou nulové vektory délky *N*). Ty jsou využity pro výpočet diskrétní fourierovy transformace (její nutnost je zdůvodněna ve vzorci (22)). Ta je vypočítána obdobně jako v *DFT2.m* tedy dvěmi vnořenými for cykly. Kde proměnné *ind* a *ind2* nabývají hodnot 1 až *N*, *pi* značí konstantu π, *1i* značí imaginární číslo. To se provede v bloku {3} V tomto výpočtu se ještě nevyužije signál zbavený své stejnosměrné složky. Ten se využije až dále.



Obr. 15: Znázornění signálů před a po S-Transformaci. (jejich rozdíl je počítán s reálnou složkou signálu po S-Transformaci, ale jak je vidět imaginární složka je zanedbatelná)



Obr. 16: Algoritmus funkce *stockwell*

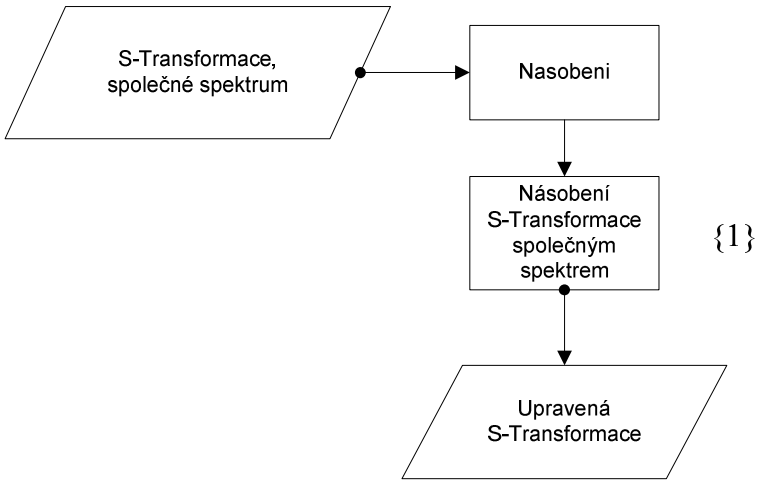
Poté se připraví nulová matice *S* (rozměru *N* x *N*) pro ukládání výsledků S-Transformace. A vektor *Q2*(délky 3 x *N*) do kterého se za „sebe“ periodicky třikrát poskládá vektor *Q* s výsledky diskrétní fourierovy transformace. To vychází ze vzorce (22) a rozsahu v něm uvedeném (*-N/2* až *N/2-1*). Samotná S-Transformace je počítána pomocí třech do sebe vnořených for cyklů. V bloku {4}, kde *ind_j*, *ind_n* a *ind* nabývají hodnot 1 až *N*, *ind_m* nabývá hodnot od *–N/2* do *N/2-1*. *e1* reprezentuje část vzorce (22) $e^{-\frac{2\pi^2 m^2}{n^2}}$ a *e2* reprezentuje $e^{\frac{i2\pi mj}{N}}$. *Q2* je jak jsem se již zmínil výše periodicky se opakující se fourierova transformace zpracovávaného signálu. Výsledná matice *S* se transponuje a uloží do proměnné *S1* (opět nulová matice rozměrů *N* x *N*). Po bloku {4} se vykreslí amplitudové spektrum spočítané S-Transformace. Pouze pro „délku“ původní sekvence. Tedy počet původních aminokyselin. Poté se z tohoto souboru volá funkce (soubor) *nasobeni.m,* který zajistí pronásobení S-Transformace se společným spektrem, blok {5}. Tato funkce bude popsána níže. Po tomto vynásobení se opět vykreslí amplitudové spektrum upravené S-Transformace násobením se společným spektrem.

Výpočet diskrétní zpětné S-Transformace proběhne v bloku {6}. A pro její výpočet jsem si zavedl další pomocné proměnné. Konkrétně *qq* a *y* kdy se jedná o nulové vektory délky *N*. Poté se pomocí třech vnořených for cyklů vypočítá diskrétní zpětná S-Transformace. Ve kterých *k, ind_n, ind_j* nabývají hodnot 1 až *N*. Kde se nejdříve pomocí pomocné proměnné *q* a skalární proměnné *R* sečte příslušný řádek z matice S-Transformace. *e3* reprezentuje $e^{\frac{i2\pi nk}{N}}$ což je část vzorce (24). A výsledný rekonstruovaný signál se uloží do

proměnné *y*. resp. z technických důvodů do proměnné *yy*. Tedy vstupem do funkce *stockwell* jsou signály, jejich délky a jejich počet spolu se společným spektrem.

5.3.1. Popis vedlejšího souboru nasobeni.m

Tento soubor (funkce) má za úkol zajistit vynásobení S-Transformace se společným spektrem. To se zajistí dvěmi vnořenými for cykly v bloku {1}.Kde *ind_1* a *ind_2* nabývají hodnot 1 až *N*. A *S1* představuje matici S-Transformace a *spektrum* reprezentuje společné spektrum. Vstupem do funkce *nasobeni* jsou tedy S-Transformace zkoumaného proteinu a společné spektrum. A výstupem je upravená S-Transformace.



Obr. 17: Algoritmus funkce *nasobeni*

5.4. Popis vedlejšího souboru energie.m

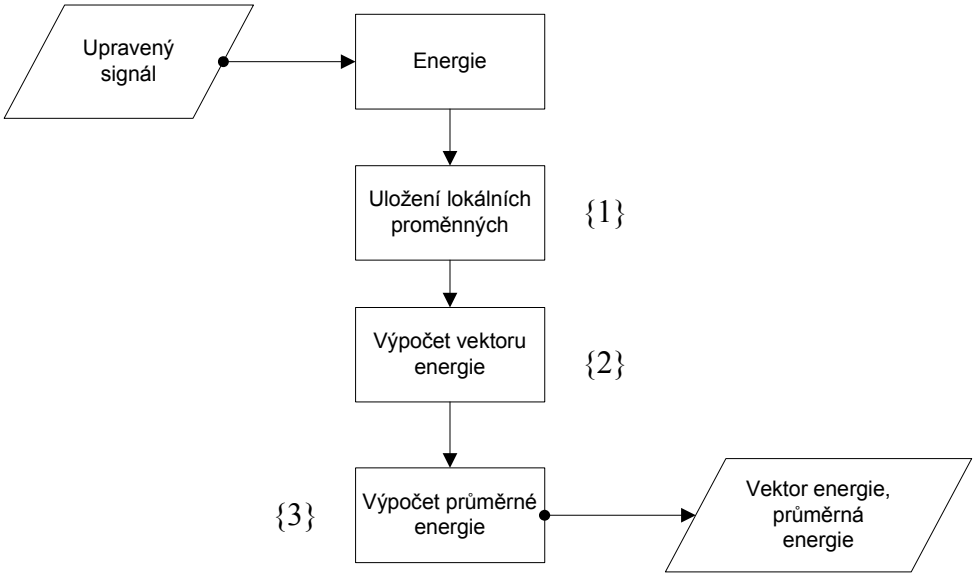
V této funkci (souboru) se provede výpočet energie jednotlivých vzorků zpracovávaného signálu i jeho průměrná energie. Energie jednotlivých vzorků se provede dle následujícího vzorce

$$E_{(n)} = \left| y_{(n)} \right|^2, \tag{17}$$

kde $\boldsymbol{E_{(n)}}$ je energie jednotlivých vzorků signálu a $\boldsymbol{y_{(n)}}$ je hodnota jednotlivých vzorků. Průměrná energie se vypočítá dle

$$E_P = \frac{1}{N} \sum_{n=1}^{N} \left| y_{(n)} \right|^2, \tag{18}$$

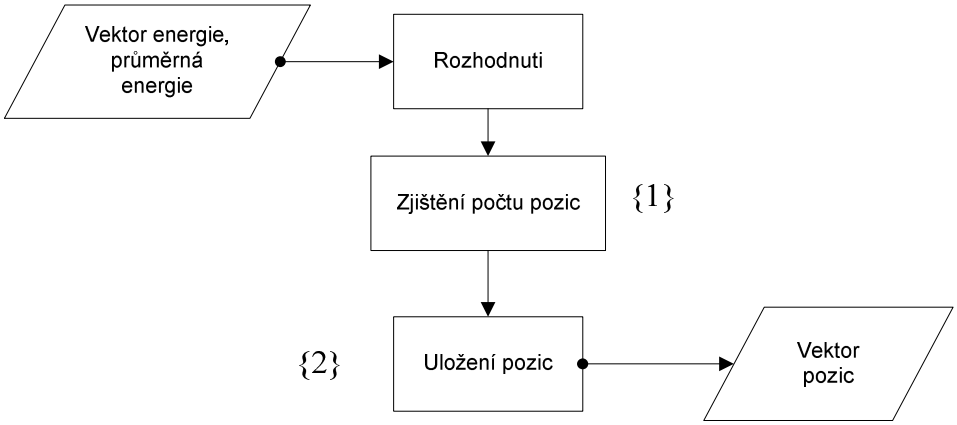
kde $\boldsymbol{E_P}$ je průměrná energie a $\boldsymbol{y_{(n)}}$ je hodnota jednotlivých vzorků. Samotný výpočet energie je realizován pomocí for cyklu v bloku {2}. Kde *N* je zde délka původní sekvence pro kterou jsem počítal S-Transformaci a *en* je nulová vektor délky *N* do kterého se jednotlivé energie ukládají. Průměrná energie je realizována pomocí následujícího řádku. Kde *prum_en* je hodnota průměrné energie signálu. Ta je počítána v bloku {3}.Vstupem do funkce energie je upravený signál z funkce *stocwell* a výstupem je vektor jeho energie a jeho průměrná energie.



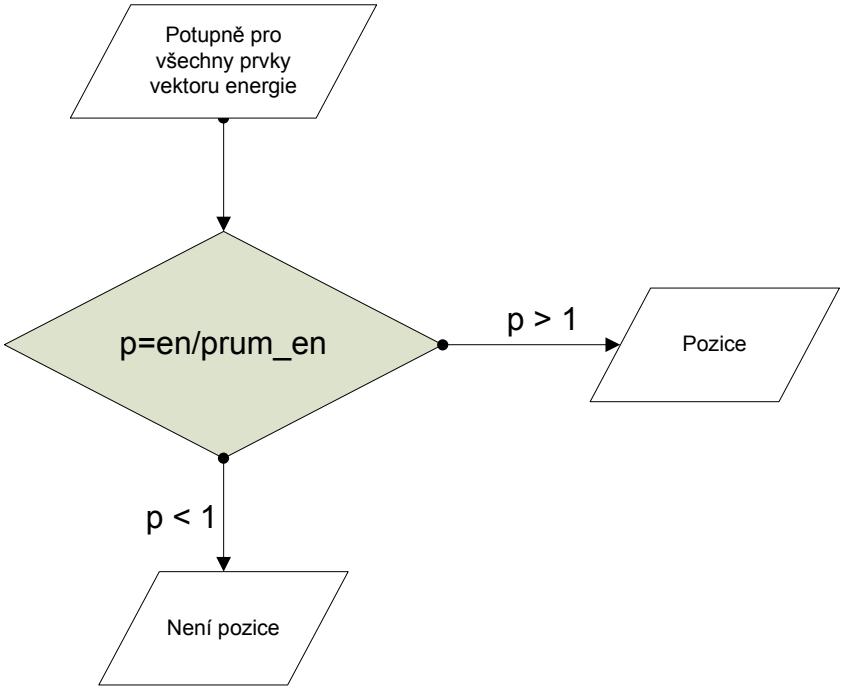
Obr. 18: Algoritmus funkce *energie*

5.5. Popis vedlejšího souboru rozhodnuti.m

Výsledkem této funkce (souboru) je vektor pozic na kterých je poměr energie jednotlivých vzorků a průměrné energie signálu větší než stanovený práh (primárně nastaven na 1). Nejdříve proběhne pomocí for cyklu s vnořeným if cyklem zjištění počtu pozic, v bloku {1}. Kde *N* je opět délka původní sekvence zpracovávané S-Transformací. V proměnné *p* se ukládá poměr energie jednotlivých vzorku a průměrné energie. Zde jsou reprezentovány proměnnými *en* (vektor jednotlivých energií vzorků signálu) a *prum_en* jako průměrná energie. Proměnná *t* je ještě před počátkem for cyklu nastavena na nulovou hodnotu a pokud je proměnná *p* větší než jedna přičte se do *t* jedna. Tím se zjistí počet pozic na kterých je poměr energie vzorku signálu na dané pozici a průměrné energie větší než jedna. Po té se vytvoří pro uložení daných pozic nulový vektor délky *t*. Pozice jsou do něj uloženy za využití podobného for cyklu s vnořeným if cyklem jako pro zjištění počtu těchto pozic, bloku {2}. Kde jsou i stejné proměnné až na to, že místo proměnné *t* je zde využita proměnná *u*. Ta je před začátkem cyklu nastavena na hodnotu rovnou jedné (jelikož Matlab indexuje vektory a matice od jedné). A samozřejmě proměnná *pozice* je nulový vektor délky *t* do kterého se ukládají jednotlivé pozice. Na obrázku č. 20 je vidět detail rozhodovacího algoritmu. Vstupem do funkce *rozhodnuti* je vektor energií pro zkoumaný signál (upravený funkcí *stockwell*) a jeho průměrná energie. Výstupem je pak vektor pozic předpokládaných aktivních míst.



Obr. 19 Algoritmus funkce *rozhodnuti*

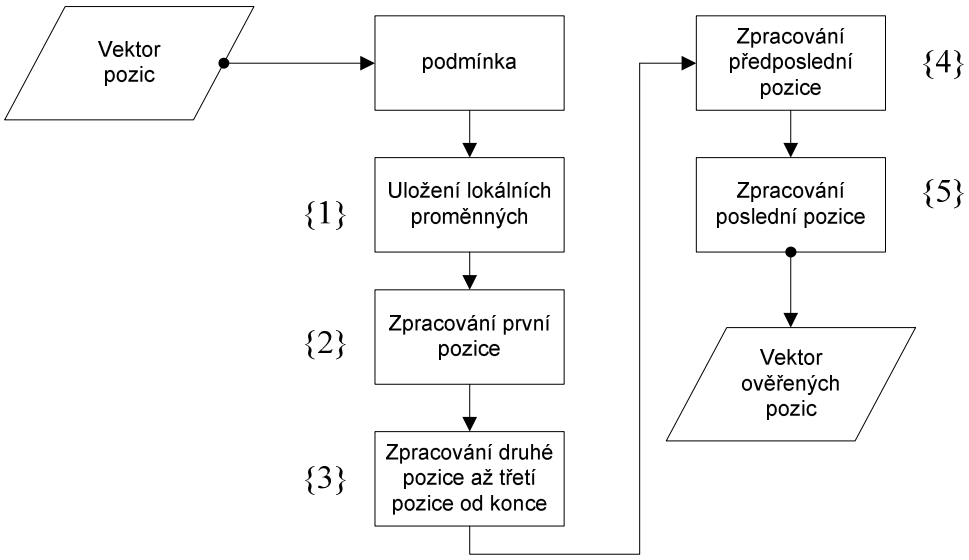


Obr. 20: Rozhodovací algoritmus z funkce *rozhodnuti*

5.6. Popis vedlejšího souboru podminka.m

Z předchozí funkce (souboru) *rozhodnuti.m* dostáváme soubor pozic na kterých je poměr energie větší než jedna. Tedy měly by se zde vyskytovat aktivní místa. Ale některé pozice jsou osamocené a je předpoklad, že aktivní místo se nemůže nalézat pouze na jedné osamocené pozici (aminokyselině). Tedy jsem zvolil podmínku,takovou aby byli označeny pouze pozice (jako aktivní místa) musí nejméně dvě ze tří sousedních pozic být určeny jako potenciální aktivní místa. K tomu jsem využil for cyklu a if cyklů. Ale vektor pozic zpracuji postupně. Nejdříve se zpracuje první pozice z vektoru pozic, v bloku {2}. Kde *pp1* a *pp2* jsou absolutní vzdálenosti mezi první pozicí a druhou pozicí, resp. třetí. A kde *b* je nulový vektor do kterého se ukládá proměnná *c* (ze začátku nastavena na hodnotu 50) vždy když je splněna podmínka pro označení jako aktivního místa. Jinak ve vektoru *b* zůstávají nulové hodnoty.

Dále se pomocí for cyklu zpracuje úsek od druhé pozice do pozice, která je dvě pozice od konce, v bloku {3}. Kde pokud absolutní vzdálenost hodnoty zkoumané pozice s hodnotami pozic o jedna větší, o dvě větší a jedna menší splňují dané podmínky tak se do proměnné *b* uloží hodnota *c*. Ta se zvětší pokud na sebe „ověřené“ pozice nenavazují. Aby bylo možno oddělit jednotlivé úseky „ověřených“ pozic. Na závěr se pomocí dvou if cyklů, obdobných jako na začátku souboru, ověří již jen dvě poslední hodnoty pozic. A proměnná *b* se uloží jako vektor ověřených pozic, bloky {4} a {5}. Vstupem do funkce *podminka* je vektor pozic a výstupem je vektor ověřených pozic.



Obr. 21: Algoritmus funkce *podminka*

6. Výsledky

V této kapitole jsou zpracovány výsledky získané mým programem a porovnané s původní prací [3] a dále se bude brát jako referenční metoda Alanin scan (ASEdb + Robetta). Provedl jsme výpočet aktivních míst pro deset proteinů (viz. Tab. 9), které jsou zpracovány i v původní práci. Byly získány ze serveru www.uniprot.org (pro společná spektra) a www.pdb.org (pro zkoumané proteiny).

Tab. 9: Zpracované proteiny

Organismus	**Název proteinu**	**č. v PDB**	**Délka sekvence**
Homo sapiens	Fibroblásotvý růstový faktor	4fgf	146
Cellulomonas fimi	Endonukleáza C	1ulo	152
Bacillus subtilis	TRP RNA-Vazebný útlumový protein	1wap	75
Homo sapiens	Lidský alpha hemoglobin	1vwt	142
Homo sapiens	Lidský růstový hormon	3hhr	190
Bacillus amyloliquefaciens	Endonukleáza	1brs	89
Bacillus amyloliquefaciens	Endonukleáza	1brs	110
Homo sapiens	Interleukin - 4	1rcb	129
Escherichia coli	Colicin E9 imunitní protein	1bxi	86
Homo sapiens	Lidský růstový hormon - vazebný protein	3hhr	203

Proteiny využité pro výpočet jednotlivých spekter jsou uvedeny v tabulkách v podkapitolách věnovaným jednotlivým proteinům.

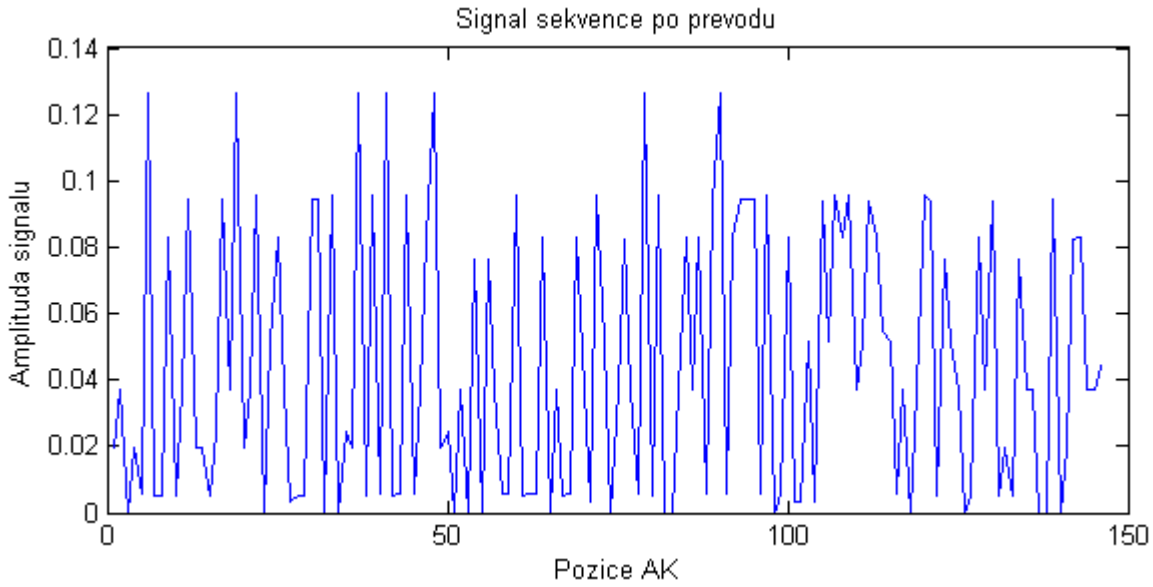
6.1. Zpracování lidského fibroblástového růstového faktoru

V tabulce 10 jsou uvedeny proteiny použité k výpočtu společného spektra. Samozřejmě společně s nimi byl použit pro výpočet společného spektra i zkoumaný protein.

Tab. 10: Proteiny použité pro výpočet společných spekter u lidského fibroblástového růstového faktoru

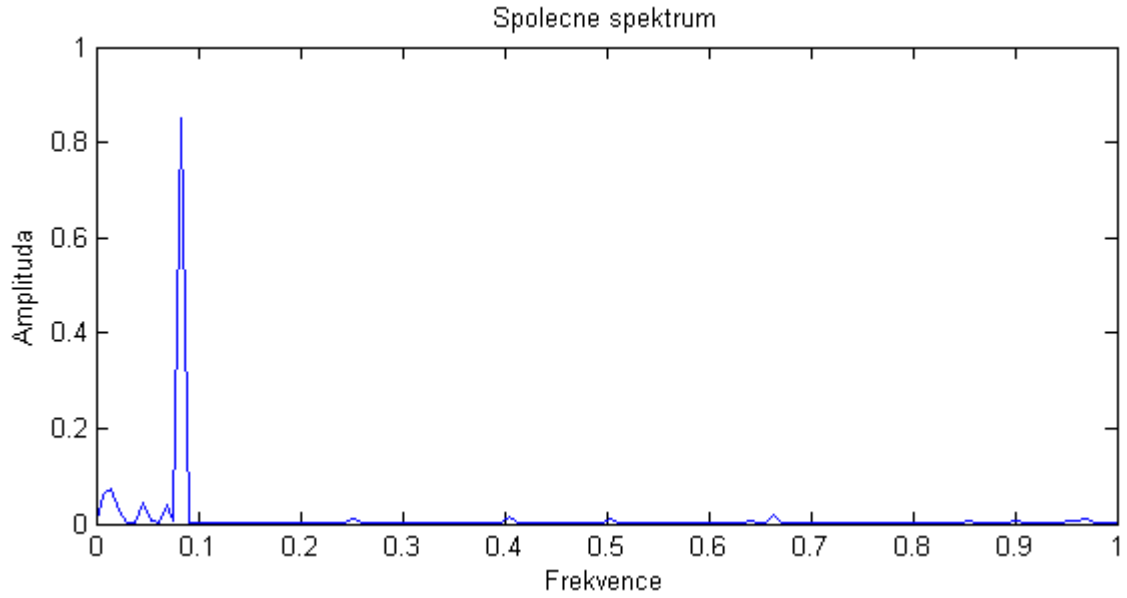
Název proteinu	**Organismus**	**Identif. Č.**
FGF1	Homo sapiens	P05230
FGF2	Homo sapiens	P09038
FGF5	Mus musculus	P15656
FGF8	Homo sapiens	P55075
FGF10	Homo sapiens	O15520
FGF16	Rattus norvegicus	O54769
FGF23	Mus musculus	Q9EPC2
FGF22	Homo sapiens	Q9HCT0
FGF-BP-1	Rattus norvegicus	Q9QY10

Jak je vidět z tabulky 10 pro výpočet společného spektra se používají i jiné než lidské proteiny (myš a krysa), které mají podobnou funkci. Na obrázku 22 je vidět zpracovávanou sekvenci, tedy její převedenou formu na signál (s využitím hodnot EIIP).

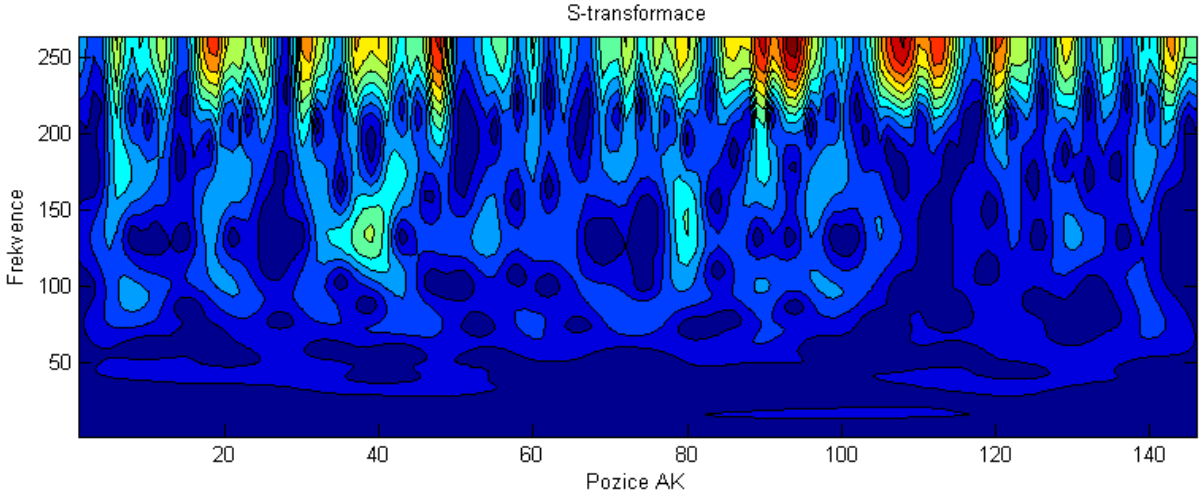


Obr. 22: Znázornění převedeného signálu (lidského fibroblásotvého faktoru) za využití hodnot EIIP

Na obrázku.23 je vidět společné spektrum proteinů z tabulky 10 a lidského fibroblástového faktoru. Jak je na něm patrné vyskytuje se dle předpokladu pouze jeden výrazný vrchol. Ten se nachází na pozici odpovídající frekvenci 0,08397 Hz. Proteiny mají tedy pouze jednu společnou funkci. Dále na obr. 24 je znázorněno amplitudové spektrum S-Transformace zkoumaného signálu (lidského fibroblásotvého faktoru), ta je ohraničena délkou zkoumané sekvence. Jelikož další hodnoty byly dodány uměle a tedy nemají pro zobrazení smysl.

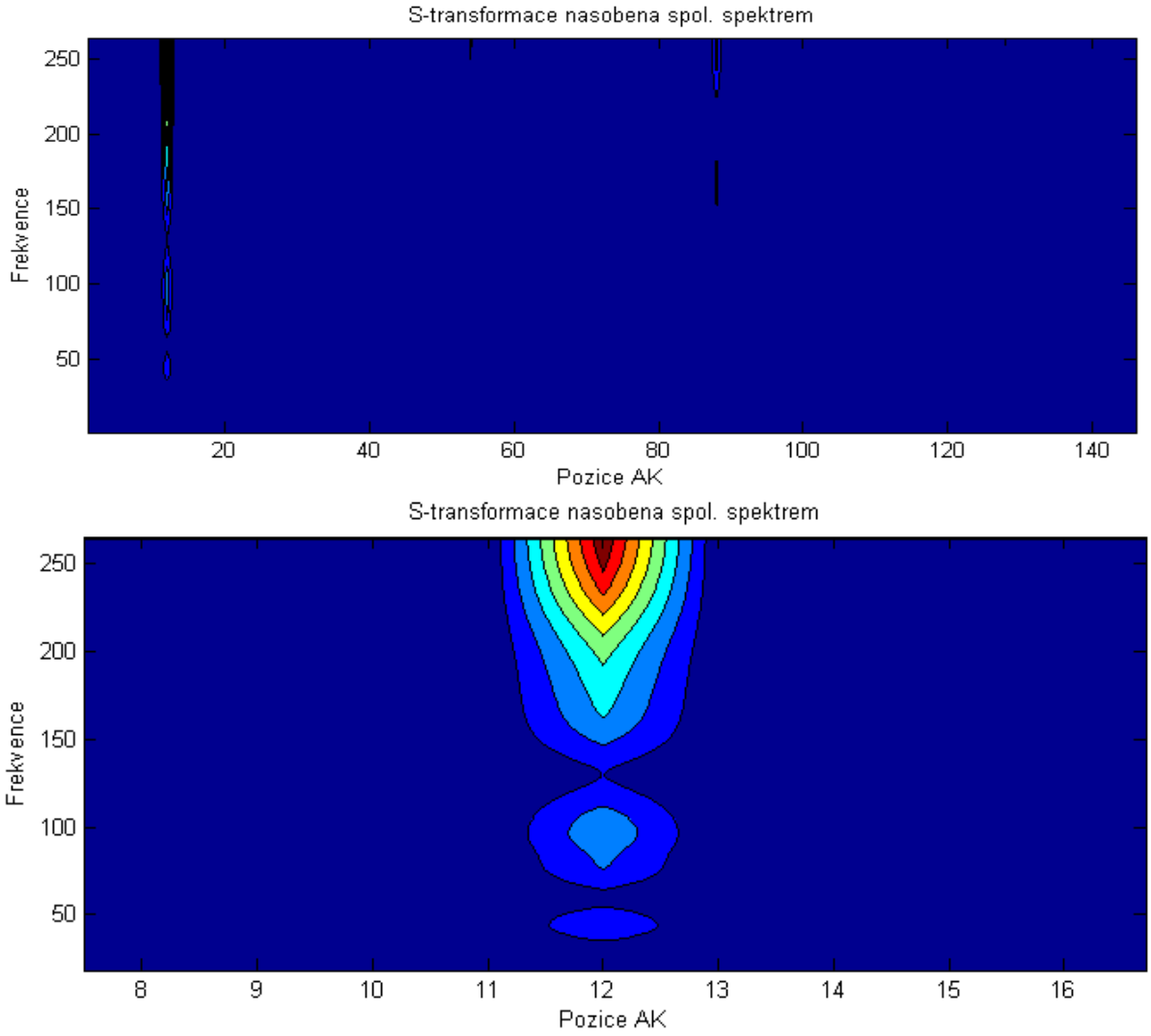


Obr. 23: Společné spektrum vypočítané z proteinů v tabulce 10 a lidského fibroblástového růstového faktoru



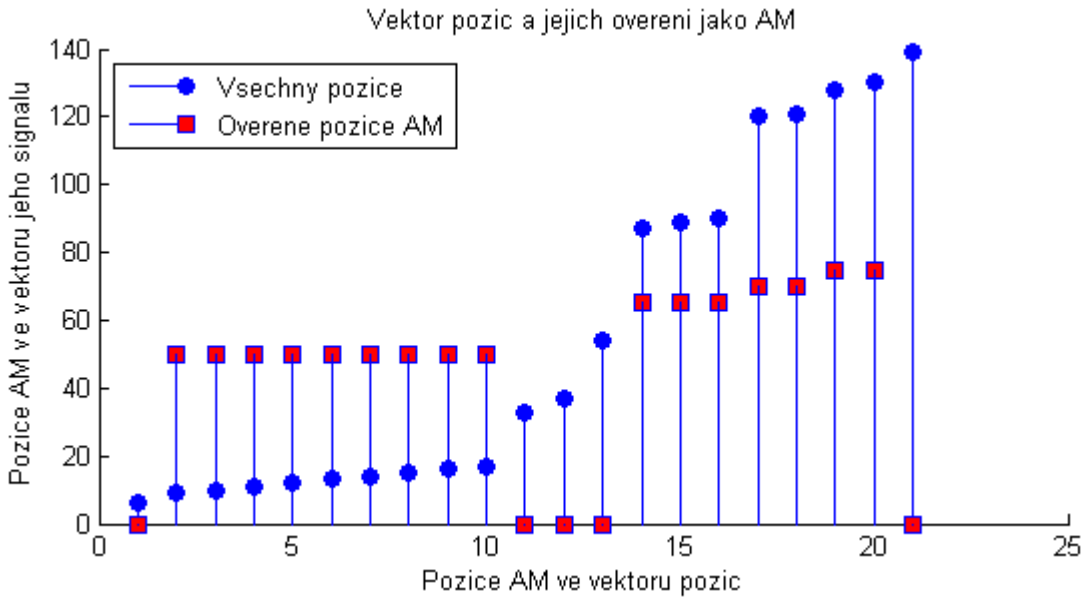
Obr. 24: Amplitudové spektrum S-Transformace zkoumaného signálu. (plná velikost obr. 24 je v přílohách)

Na obrázku 25 vidět znázornění S-Transformace vynásobené společným spektrum, resp. opět amplitudové spektrum.

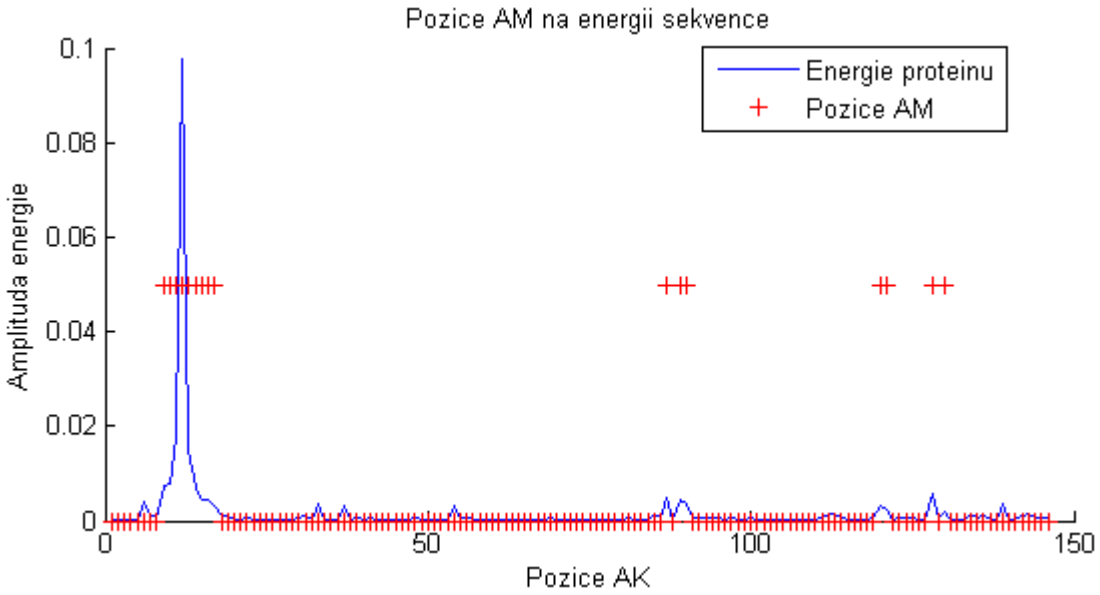


Obr. 25: Amplitudové spektrum S-Transformace násobené společným spektrem (nahoře – celé, dole – detail vrcholu okolo 12 pozice přes všechny frekvence)

Po zpětné transformaci dostáváme upravený signál (násobením se společným spektrem), ze kterého se určí energie a následně i pozice aktivních míst. To je vidět na obrázcích 26 a 27.



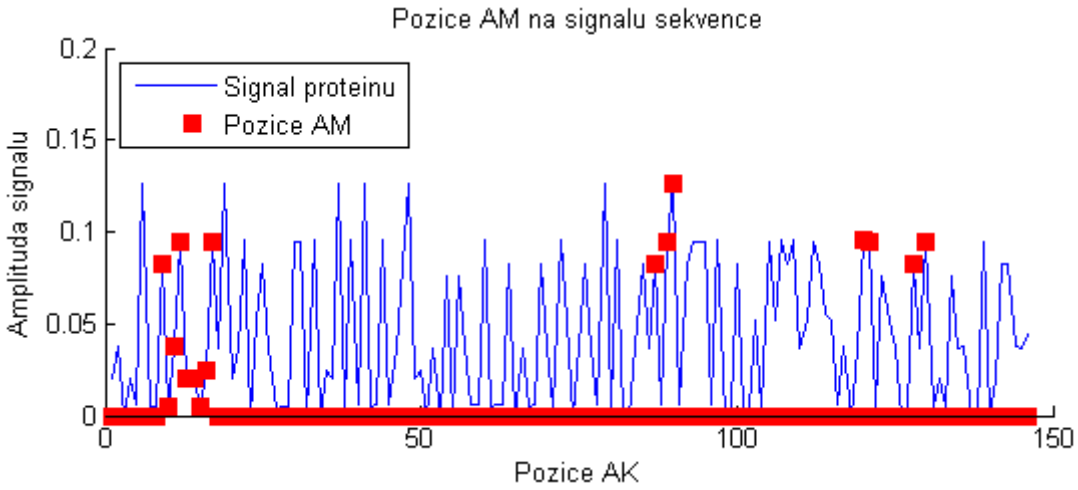
Obr. 26: Vektory pozic a jejich ověření.



Obr. č. 27: Vektory energie a pozice ověřených AM

Pozice určené mým programem tedy jsou čtyři skupiny aminokyselin. První se nachází na pozicích: 9, 10, 11, 12, 13 ,14, 15, 16, 17. Což ve fasta kódu odpovídá sekvenci znaků SGAFPPGHF. Ta reprezentuje následující sekvenci aminokyselin: Ser-Gly-Ala-Phe-Pro-Pro-Gly-His-Phe. Další skupina se nachází na pozicích: 87, 89, 90. Ty ve fasta kódu odovídají sekvenci znaků C-TD. A ta reprezentuje následující aminokyseliny: Cys-GAP-Thr-Asp. Třetí skupina se nachází na pozicích: 120 a 121. Tedy ve fasta kódu znaky: RT. A ty odpovídají aminokyselinám: Arg-Thr. A poslední skupina se nachází na pozicích: 128, 130. Ty ve fasta kódu reprezentují znaky S-T, což představuje aminokyseliny

Ser-GAP-Thr. Na obrázku 28 je vidět pozice aktivních míst zobrazených na signálu zkoumané sekvence.



Obr. 28: Signál původní sekvence a pozice ověřených AM

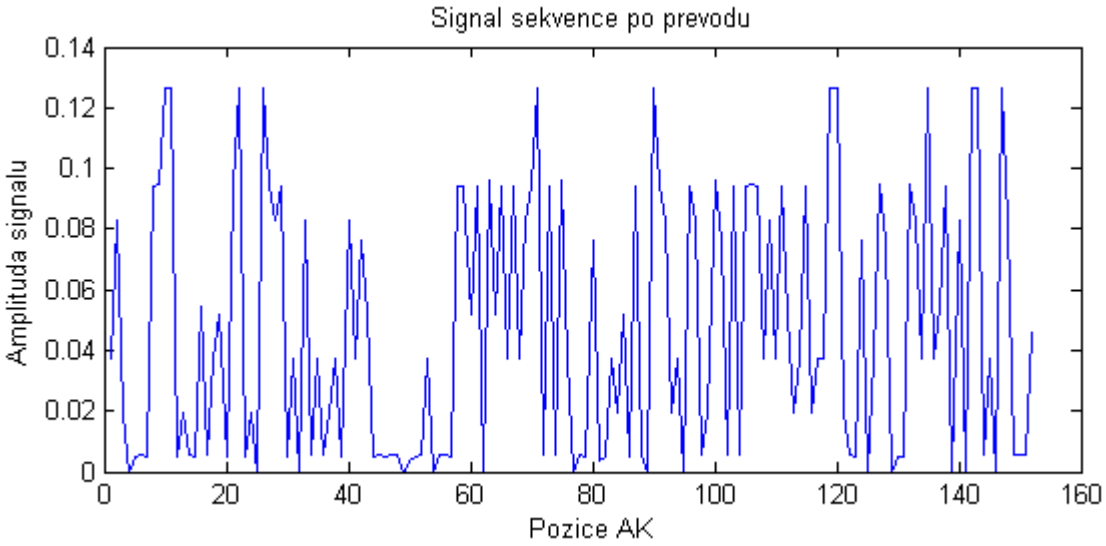
6.2. Zpracování endonukleázy C z Cellulomonas fimi

V tabulce 11 jsou uvedeny proteiny použité k výpočtu společného spektra. Společně s nimi byl použit pro výpočet společného spektra i zkoumaný protein (endonukleáza C z Cellulomonas fimi).

Tab. 11: Proteiny použité pro výpočet společného spektra

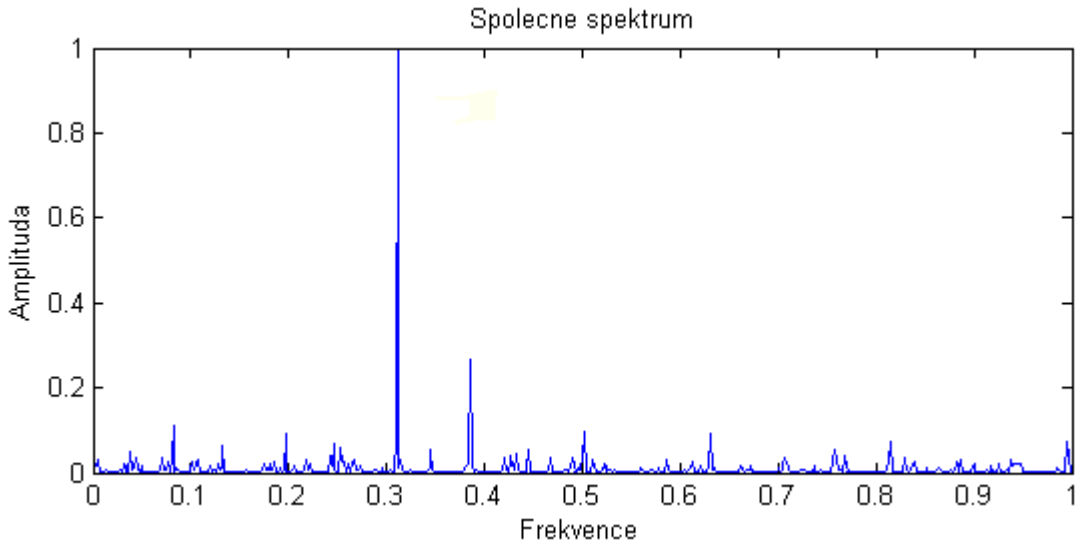
Název proteinu	Organismus	Identif. Č.
celC	Clostridium thermocellum	P0C2S3
cenC	Cellulomonas fimi	P14090
celC	Bacillus sp.	P19570
celC	Cellvibrio japonicus	P27033
celCCC	Clostridium cellulolyticum	P37699
Endoglucanase CX	Prunus persica	P38534
Endoglucanase	Ralstonia solanacearum	Q93GB3
celC	Clostridium thermocellum	A3DJ77

Jak je vidět u tohoto společného spektra se více využívají podobné proteiny ale z jiných organismů. Na obrázku 29 je vidět signál zpracovávané sekvence.

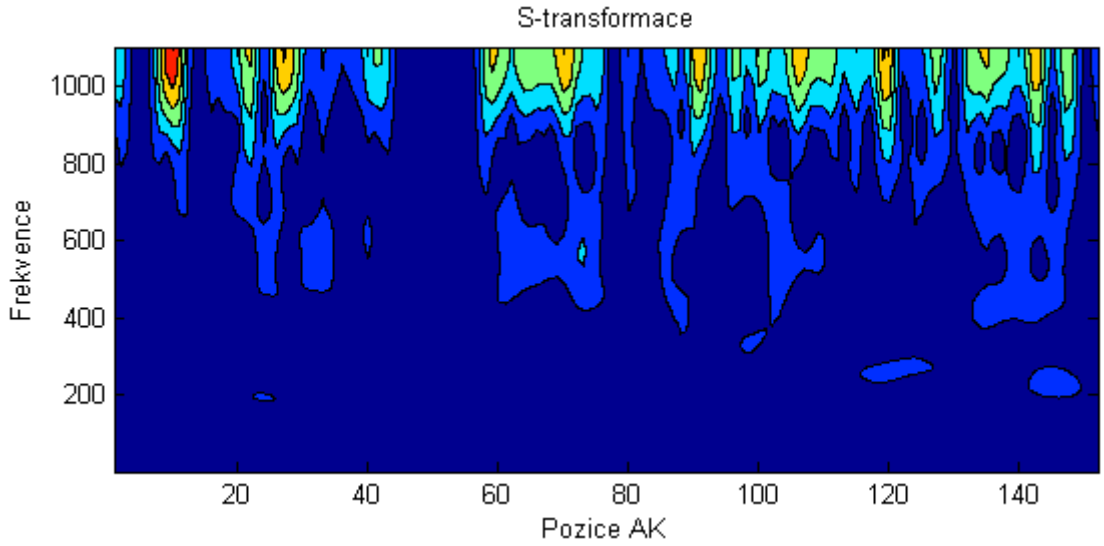


Obr. 29: Znázornění převedeného signálu (endonukléaza C z Cellulomonas fimi) za využití hodnot EIIP

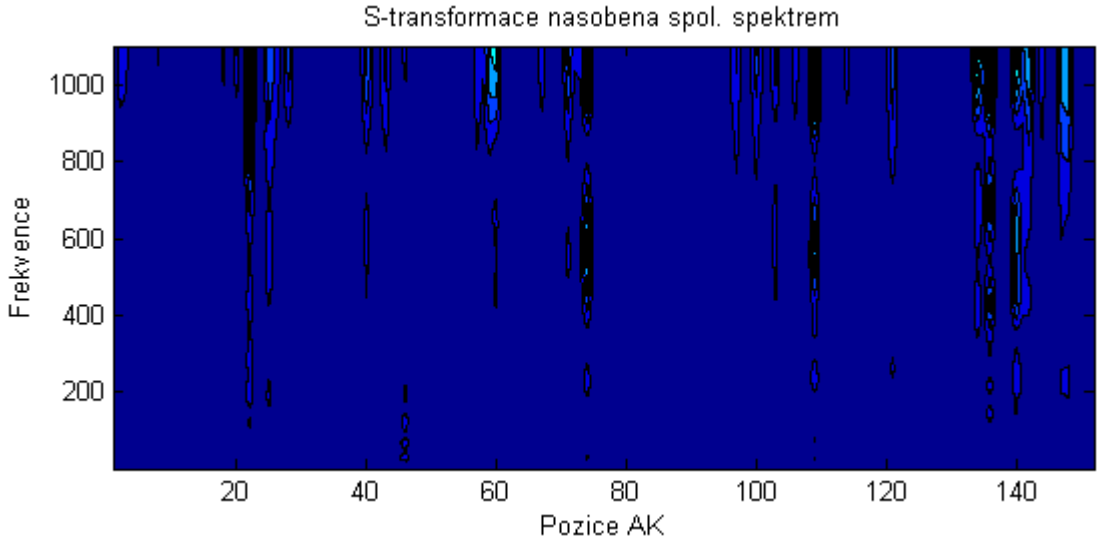
Na obrázku 30 je vidět společné spektrum proteinů z tabulky 11 a zpracovávaného proteinu (zobrazeného na obr. 29). Na něm je patrný jeden velmi výrazný vrchol a jeden menší ale ten má pouze asi pětinovou výšku nejvyššího. Tento nejvyšší vrchol se nachází na frekvenci 0,3109 Hz. Proteiny tedy mají opět jednu společnou funkci. Na obrázku 31 je znázorněno amplitudové spektrum S-Transformace zkoumaného proteinu. Ta je opět ohraničena jeho délkou jelikož další hodnoty byli dodány uměly a tedy nemají smysl pro zobrazení. Na obrázku 32 je vidět amplitudové spektrum S-Transformace po vynásobení se společným spektrem (plná velikost obr. 32 je v přílohách).



Obr. 30: Společné spektrum vypočítané z proteinů v tabulce 11 a endonukleázy C z Cellulomonas fimi

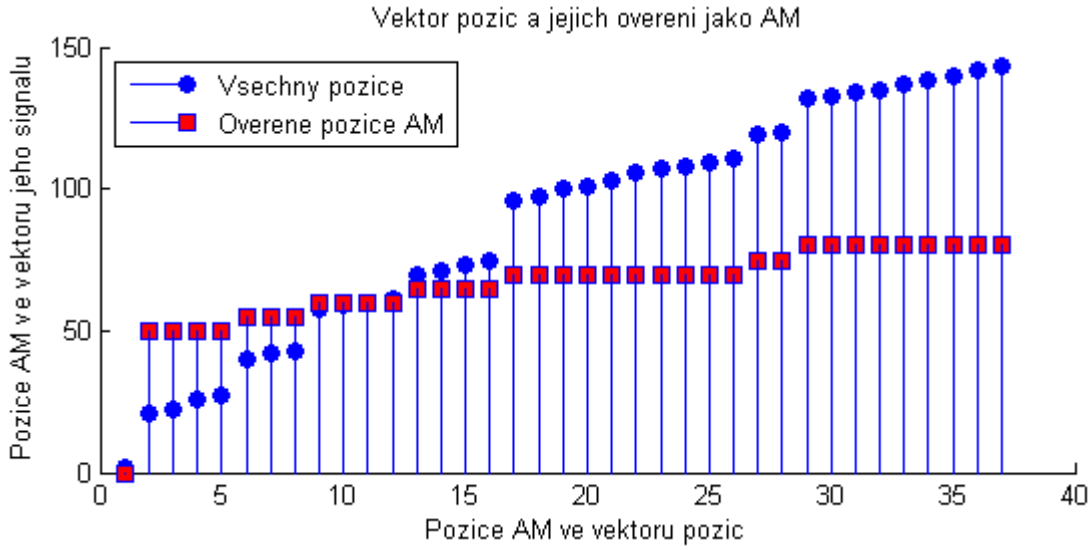


Obr. 31: Amplitudové spektrum S-Transformace zkoumaného signálu

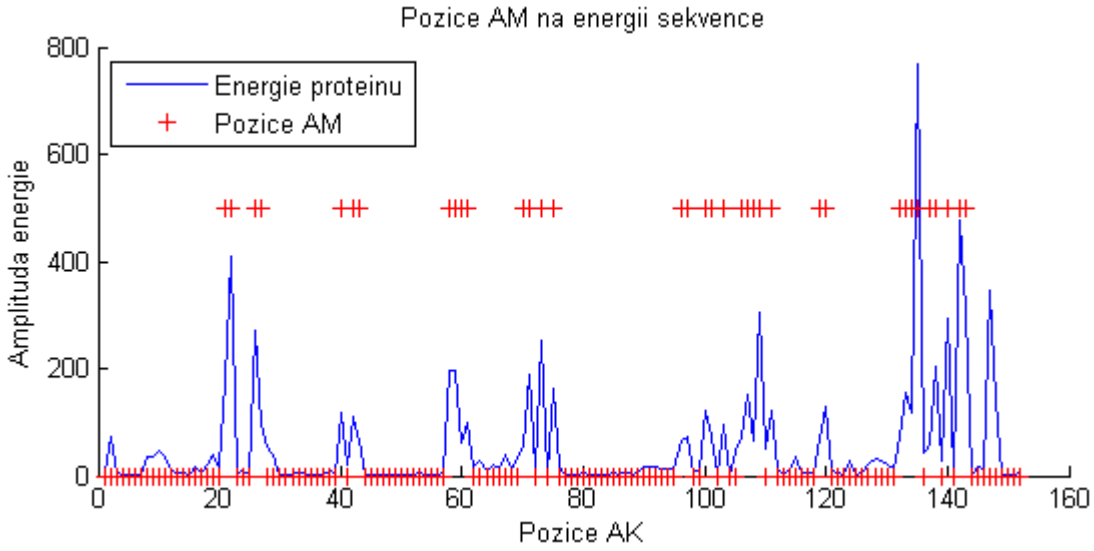


Obr. 32: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem

Po zpětné transformaci dostáváme upravený signál (násobením se společným spektrem), ze kterého se určí energie a následně i pozice aktivních míst. To je vidět na obrázcích 33 a 34. A jak vyplívá z obrázku č.:26 v tomto případě skoro všechny pozice se ověřily jako aktivní místa. Jak je dále také patrné u tohoto zkoumaného proteinu je osm skupin označených jako aktivní místa. První se nachází na pozicích 21 a 22. Ty ve fasta kódu sekvence jsou reprezentovány znaky TD a ty reprezentují aminokyseliny Thr-Asp. Druhá skupina aminokyselin se nachází na pozicích 26 a 27. Což ve fasta kódu sekvence znamená znaky DT a ty reprezentují následující aminokyseliny Asp-Thr. Třetí skupina se rozkládá na pozicích 40, 42, 43. A ty ve fasta kódu sekvence ukazují na znaky S-QY neboli aminokyseliny Ser-GAP-Gln-Tyr. Čtvrtá skupina je na pozicích 58, 59, 60, 61. Neboli ve fasta kódu sekvence TTYT což představuje následující aminokyseliny Thr-Thr-Tyr-Thr. Pátá skupina aminokyselin se nachází na pozicích 70, 71, 73, 75 neboli ve fasta kódu sekvence znamená TD-T-R a tento sled znaků ukrývá aminokyseliny Thr-Asp-GAP-Thr-GAP-Arg.

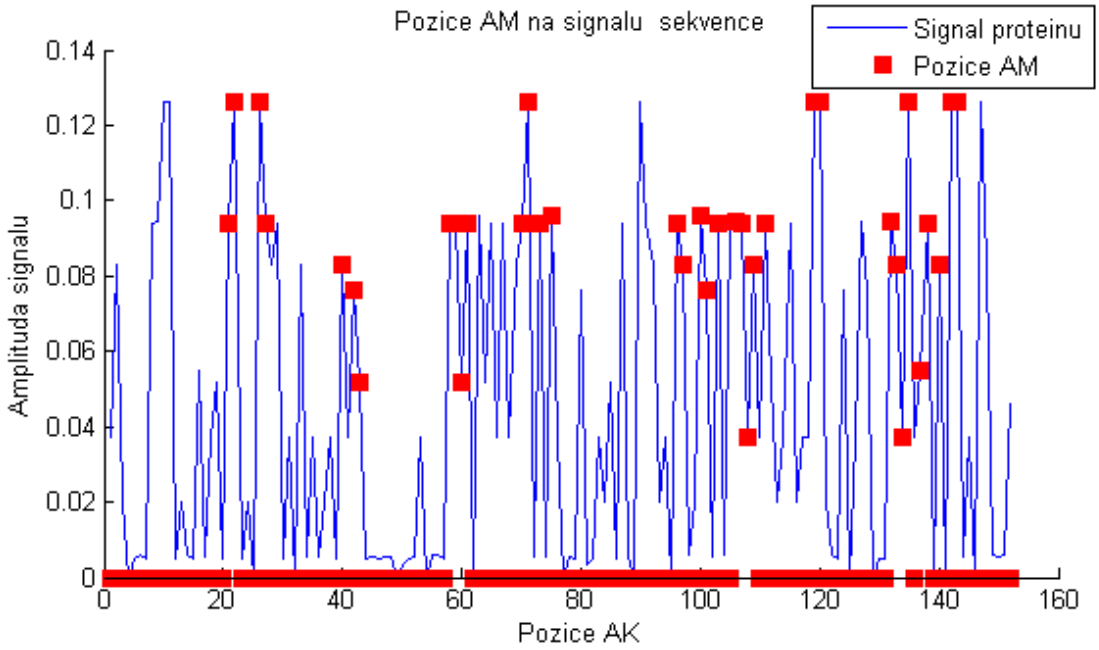


Obr. 33: Vektory pozic a jejich ověření



Obr. 34: Vektory energie a pozice ověřených AM

Šestá skupina pozic je na 96, 97, 100, 101, 103, 106, 107, 108, 109, 111 ty reprezentují ve fasta kódu sekvence následující řetězec TS--RQ-T--FTAS-T neboli aminokyseliny Thr-Ser-GAP-GAP-Arg-Gln-GAP-Thr-GAP-GAP-Phe-Thr-Ala-Ser-GA-PThr. Sedmá skupina pozic se nachází na pozicích 119 a 120 a ty ve fasta kódu představují znaky DD což představuje aminokyseliny Asp-Asp. A poslední osmá skupina se nachází na pozicích 132, 133, 134, 135, 137, 138, 140, 142, 143 a ty ve fasta kódu reprezentují řetězec FSAD-WT-C-DD neboli aminokyseliny Phe-Ser-Ala-Asp-GAP-Trp-Thr-GAP-Cys-GAP-Asp-Asp. A na obrázku 35 je vidět signál původní sekvence s vyznačenými aktivními místy.



Obr. 35: Signál původní sekvence a pozice ověřených AM

6.3. Zpracování TRP RNA-Vazebného útlumového proteinu z Bacillus subtilis

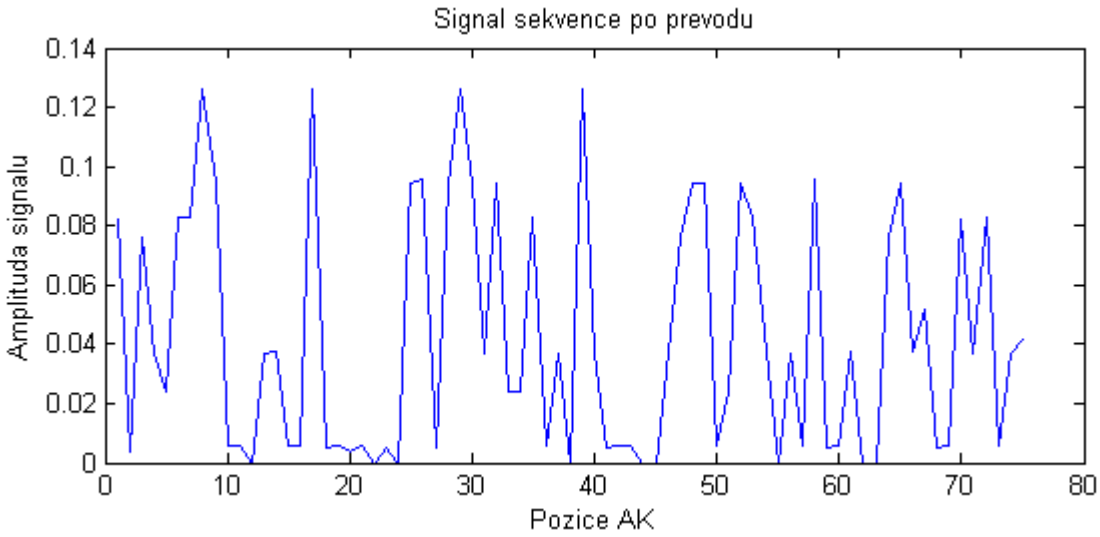
V tabulce 12 jsou uvedeny proteiny použité k výpočtu společného spektra. Společně s nimi byl použit pro výpočet společného spektra i zkoumaný protein (TRP RNA-vazebný protein z bacillus subtilis).

Tab. 12: Proteiny použité pro výpočet společného spektra

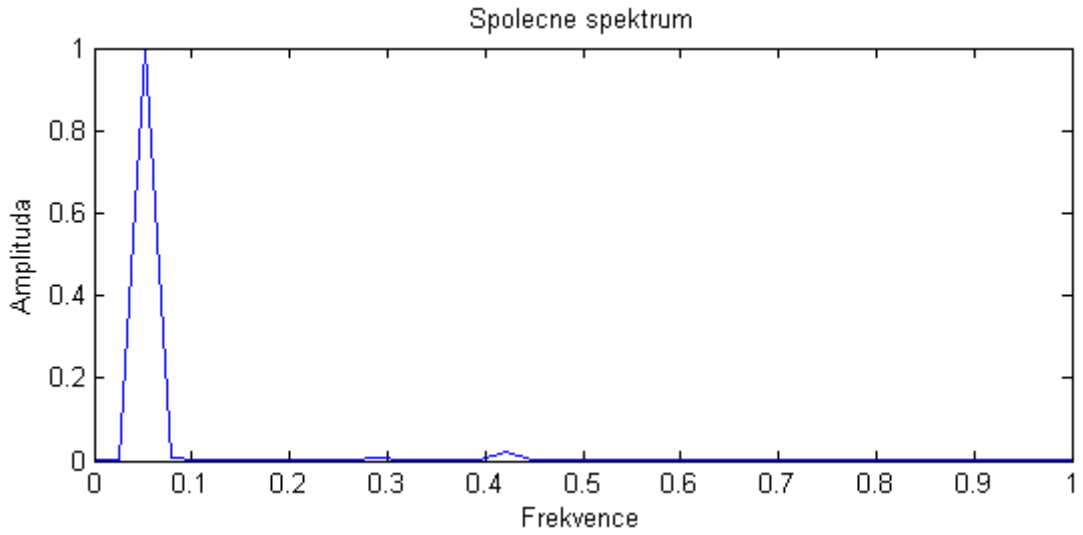
Název proteinu	**Organismus**	**Identif. Č.**
Transkripční útlumový protein mtrB	Bacillus subtilis	P19466
Transkripční útlumový protein mtrB	Bacillus pumilus	P48064
Tryptophan RNA-vazebný útlumový protein	Moorella thermoacetica	Q2RHB9
Transkripční útlumový protein mtrB	Oceanobacillus iheyensis	Q8EQB3
Transkripční útlumový protein mtrB	Geobacillus sp.	C5D3E7
Transkripční útlumový protein mtrB	Bacillus stearothermophilus	Q9X6J6

Pro výpočet tohoto společného spektra jsou opět použity většinou podobné proteiny od různých organismů. Signál zkoumaného proteinu je vidět na obrázku 36. Na obrázku 37 je vidět společné spektrum proteinů z tabulky 12 a zpracovávaného proteinu (zobrazeného na obrázku 36). Na něm je patrný jeden velmi výrazný vrchol. Tento vrchol se nachází na frekvenci 0,05263 Hz. Jeho „plochost“ je patrně způsobena délkou zpracovávaných sekvencí. V tomto případě jsou to velmi krátké sekvence (kratší než 100 aminokyselin). Proteiny mají opět jednu společnou funkci. Na obrázku 38 je znázorněno amplitudové spektrum

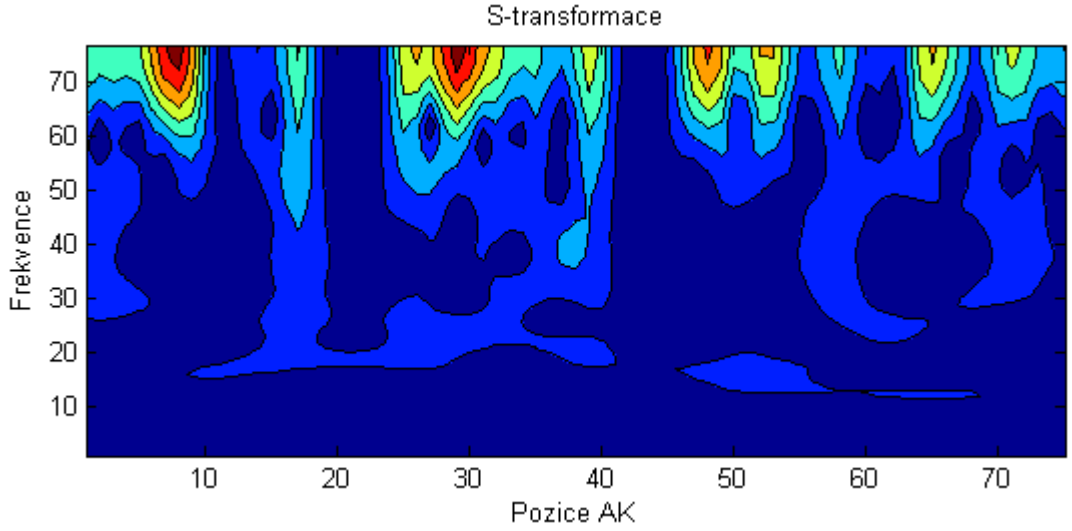
S-Transformace zkoumaného proteinu. Ta je opět ohraničena jeho délkou jelikož další hodnoty byly dodány uměly a tedy nemají smysl pro zobrazení. Na obrázku 39 je vidět amplitudové spektrum S-Transformace po vynásobení se společným spektrem (plná velikost obr. 39 je v přílohách).



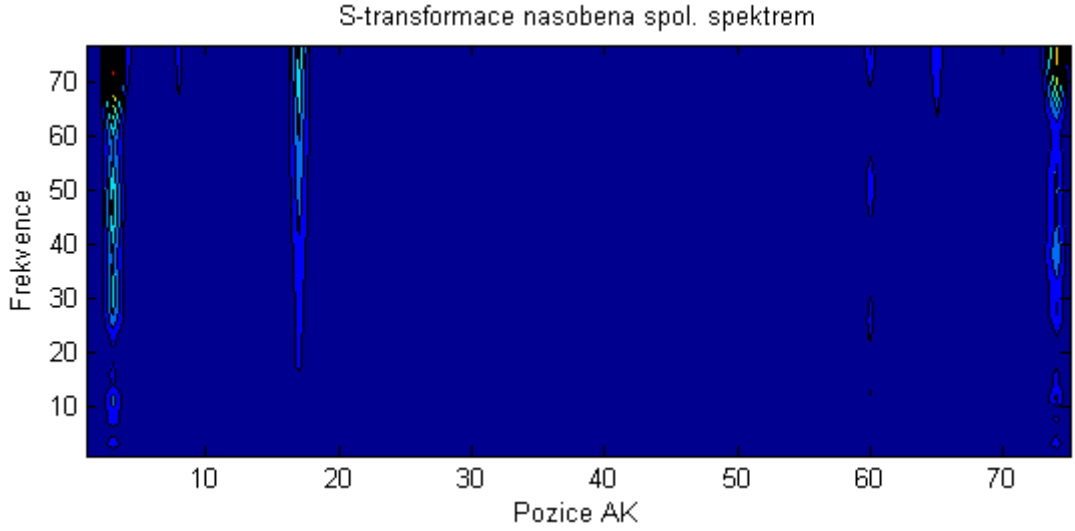
Obr. 36: Znázornění převedeného signálu (TRP RNA-vazebný protein z bacillus subtilis) za využití hodnot EIIP



Obr. 37: Společné spektrum vypočítané z proteinů v tabulce 12 a TRP RNA-vazebného proteinu z bacillus subtilit

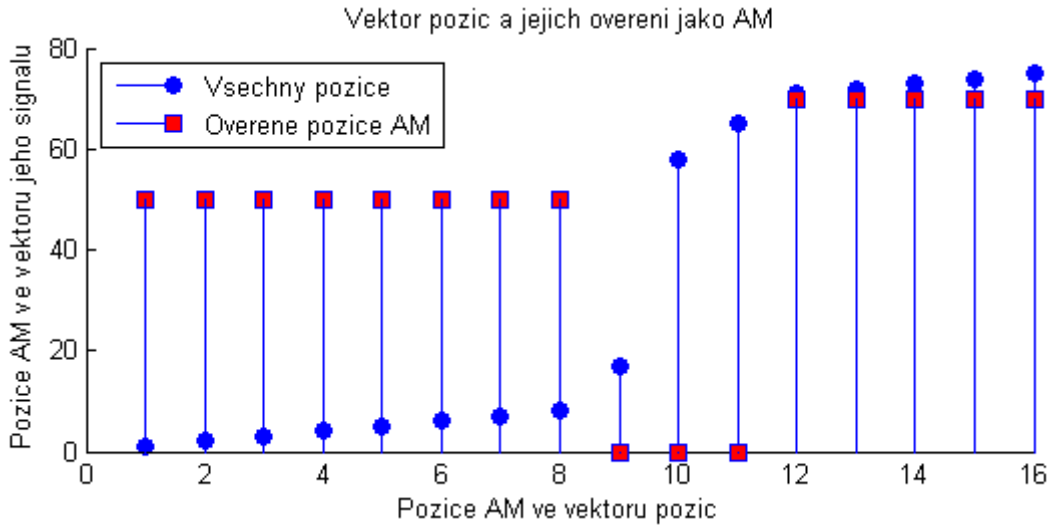


Obr. 38: Amplitudové spektrum S-Transformace zkoumaného signálu

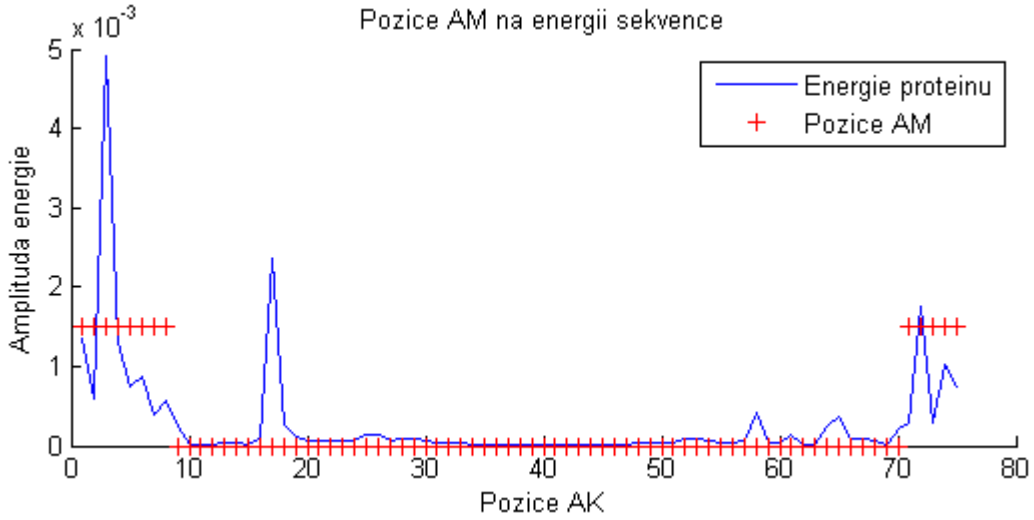


Obr. 39: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem

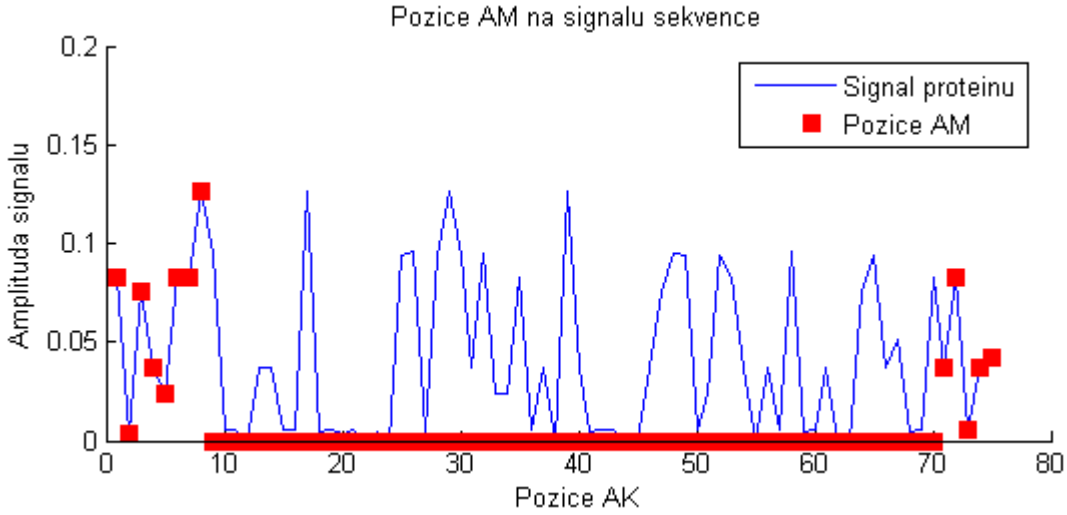
Po zpětné transformaci dostáváme upravený signál (násobením se společným spektrem), ze kterého se určí energie a následně i pozice aktivních míst. To je vidět na obrázcích 40 a 41. Jak je vidět z obrázku 40 tak v tomto případě dostáváme dvě skupiny aminokyselin (vždy na koncích sekvencí). Tedy první skupina se nachází na pozicích 1, 2, 3, 4, 5, 6, 7, 8 což ve fasta kódu sekvence znamená řetězec znaků MNQKHSSD. Ty reprezentují následující aminokyseliny Met-Asn-Gln-Lys-His-Ser-Ser-Asp. Druhá skupina se nachází na pozicích 71, 72, 73, 74, 75 ty ve fasta kódu sekvence představují znaky KSEKK. Ty reprezentují tyto aminokyseliny Lys-Ser-Glu-Lys-Lys. Na obrázku 42 je vidět sekvence původního signálu s vyznačenými aktivními místy.



Obr. 40: Vektory pozic a jejich ověření



Obr. 41: Vektory energie a pozice ověřených AM



Obr. 42: Signál původní sekvence a pozice ověřených AM

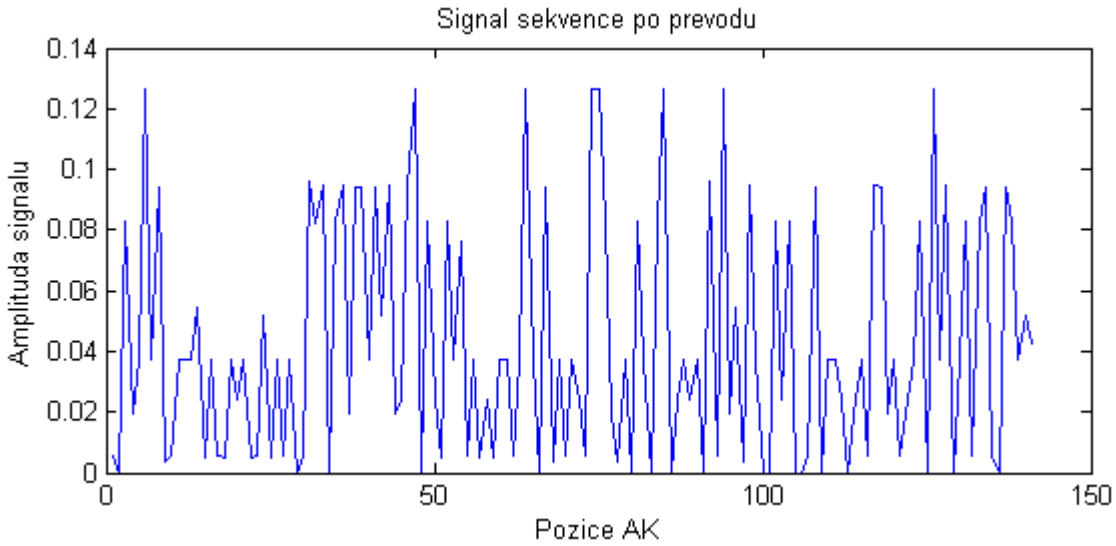
6.4. Zpracování lidského alpha hemoglobinu

V tabulce 13 jsou uvedeny proteiny použité k výpočtu společného spektra. Samozřejmě společně s nimi byl použit pro výpočet společného spektra i zkoumaný protein (lidský alpha hemoglobin).

Tab. 13: Proteiny použité pro výpočet společného spektra

Název proteinu	**Organismus**	**Identif. Č.**
Hemoglobin podjednotka beta	Canis familiaris	P60524
Hemoglobin podjednotka alfa	Equus caballus	P01958
Hemoglobin podjednotka beta	Equus caballus	P02062
Hemoglobin podjednotka alfa	Homo sapiens	P69905
Hemoglobin podjednotka beta	Homo sapiens	P68871
Hemoglobin podjednotka beta-1	Panthera leo	P68050
Hemoglobin podjednotka alfa	Mus musculus	P01942
Hemoglobin podjednotka alfa-1/2	Rattus norvegicus	P01946
Hemoglobin podjednotka beta-1	Panthera tigris sumatrae	P68048

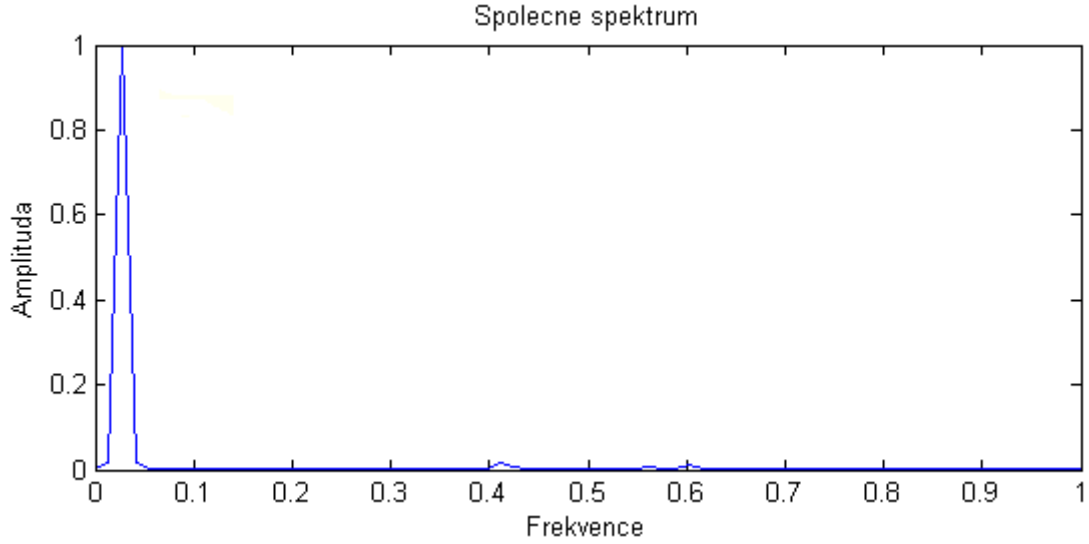
Jak je vidět jsou opět využity proteiny z různých organismů (v tomto případě zvířat). A to jak podjednotky alfa tak i podjednotky beta. Signál zkoumaného proteinu můžeme vidět na obrázku 43, tedy lidského hemoglobinu podjednotky alfa.



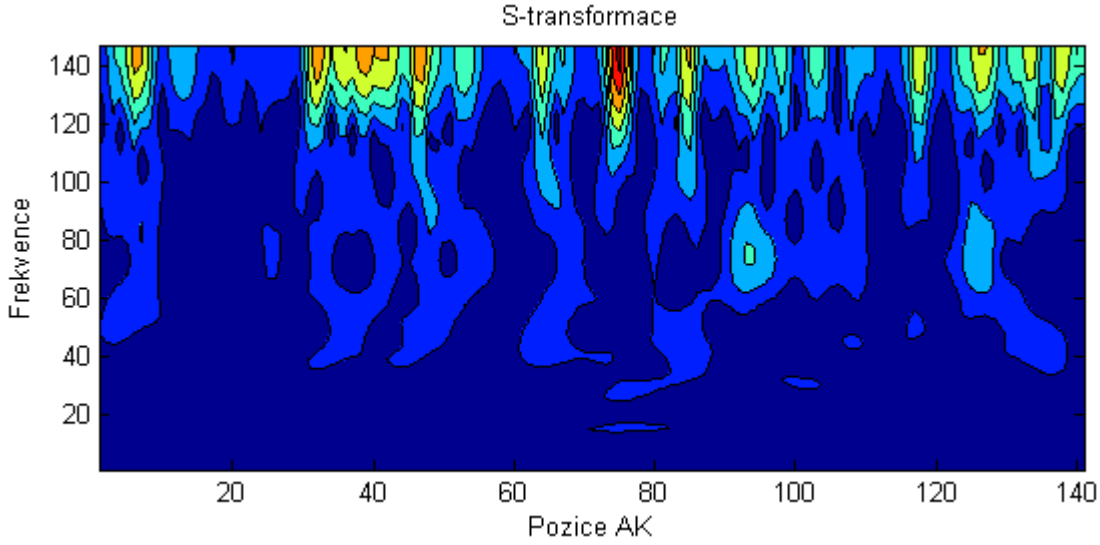
Obr. 43: Znázornění převedeného signálu (lidský hemoglobin podjednotka alfa) za využití hodnot EIIP

Na obrázku 44 je vidět společné spektrum proteinů z tabulky 13 a zpracovávaného proteinu (zobrazeného na obr. 43). Na něm je patrný jeden velmi výrazný vrchol. Tento vrchol se nachází na frekvenci 0,0274 Hz. Proteiny mají opět jedinou společnou funkci. Na obrázku 45

je znázorněno amplitudové spektrum S-Transformace zkoumaného proteinu. Ta je opět ohraničena jeho délkou jelikož další hodnoty byli dodány uměly a tedy nemají smysl pro zobrazení. Na obrázku 46 je vidět amplitudové spektrum S-Transformace po vynásobení se společným spektrem (plná velikost obr. 46 je v přílohách).



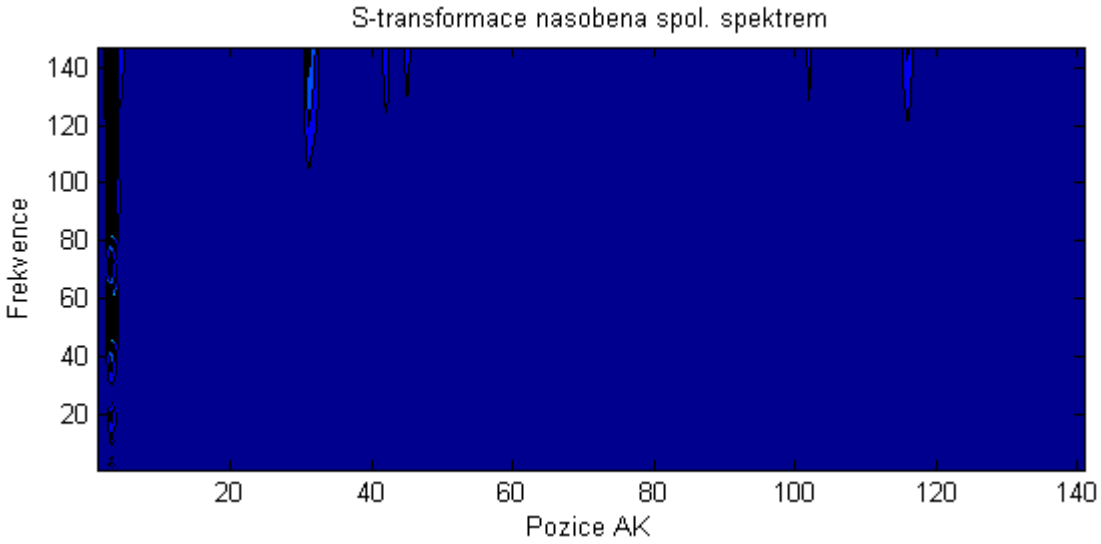
Obr. 44: Společné spektrum vypočítané z proteinů v tabulce 13 a lidského hemoglobinu podjednotky alfa



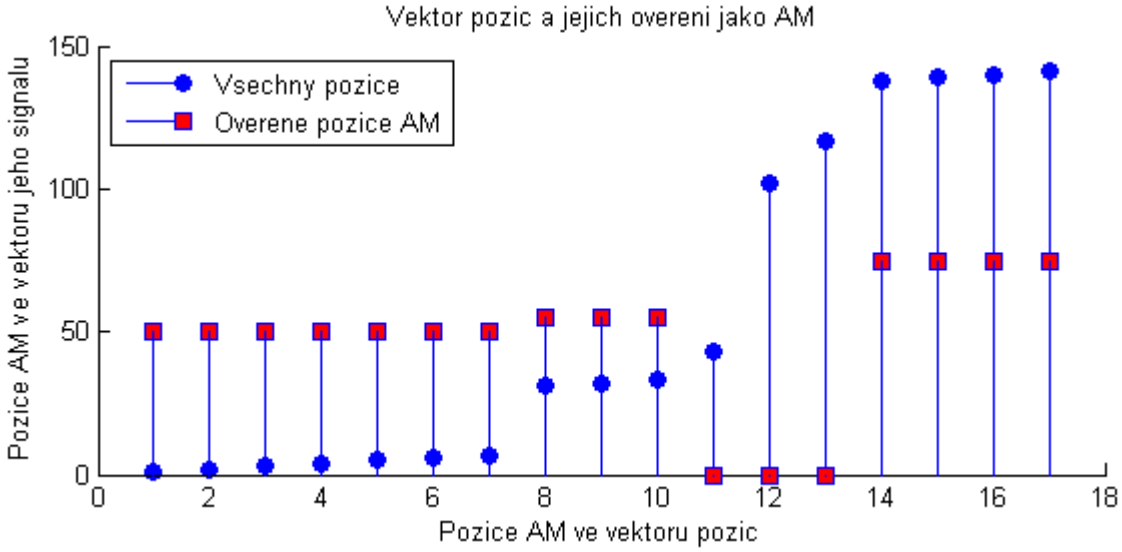
Obr. 45: Amplitudové spektrum S-Transformace zkoumaného signálu

Po zpětné transformaci dostáváme upravený signál (násobením se společným spektrem), ze kterého se určí energie a následně i pozice aktivních míst. To je vidět na obrázcích 47 a 48. Jak je vidět z obrázku 47 tak v tomto případě dostáváme tři skupiny aminokyselin (dvě opět na koncích sekvencí). Prní skupina se nachází na pozicích 1, 2, 3, 4, 5, 6, 7 což ve fasta kódu sekvence ukrývá znaky VLSPADK a ty představují aminokyseliny Val-Leu-Ser-Pro-Ala-Asp-Lys. Druhá skupina se nachází na pozicích 31, 32, 33 ukrývajících ve fasta kódu sekvence znaky RMF. Ty představují aminokyseliny Arg-Met-Phe. A poslední skupina pozic je na 138,

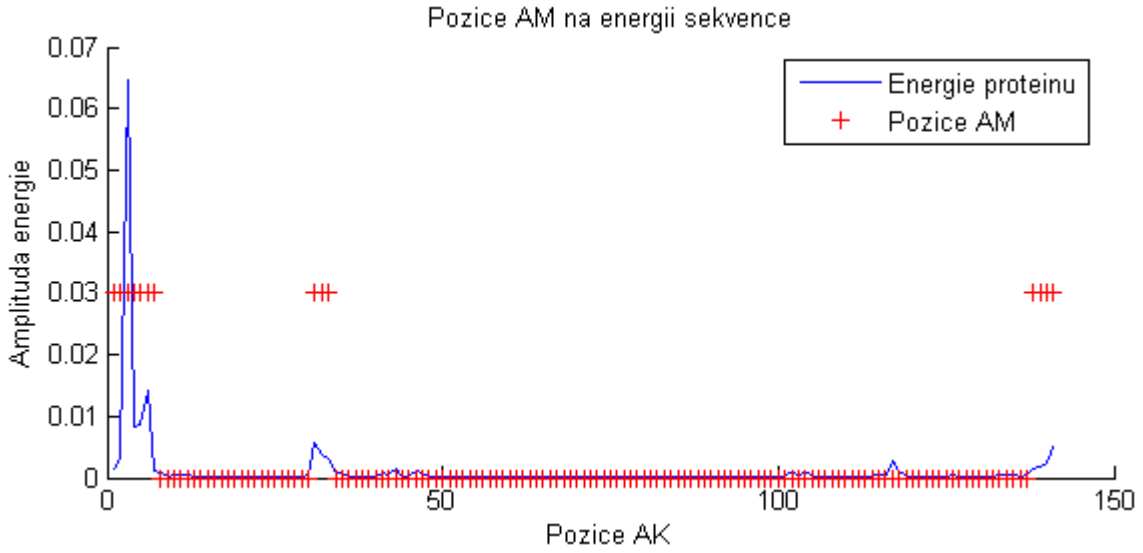
139, 140, 141 ty ve fasta kódu sekvence představují znaky SKYR. Ty reprezentují aminokyseliny Ser-Lys-Tyr-Arg. Na obrázku 49 je vidět sekvence původního signálu s vyznačenými aktivními místy.



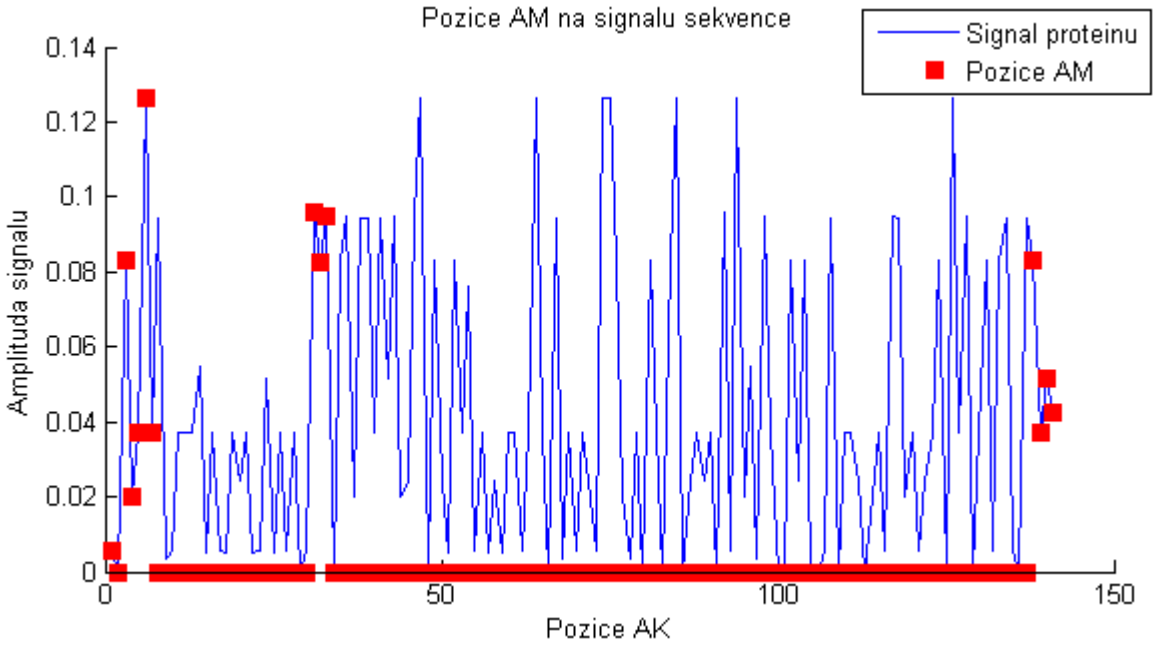
Obr. 46: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem



Obr. 47: Vektory pozic a jejich ověření



Obr. 48: Vektor energie a pozice ověřených AM



Obr. 49: Signál původní sekvence a pozice ověřených AM

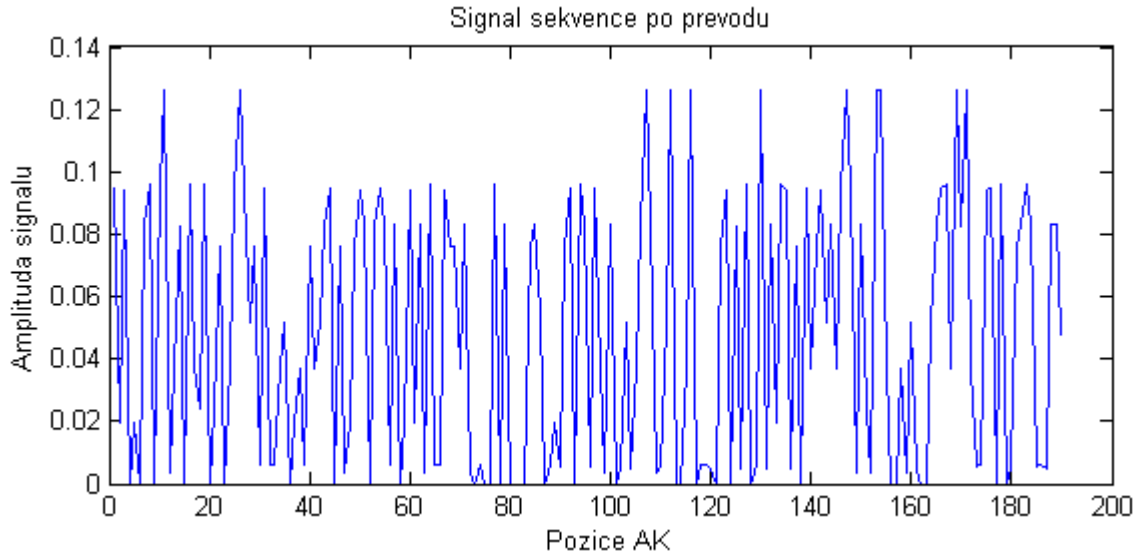
6.5. Zpracování lidského růstového hormonu

V tabulce 14 jsou uvedeny proteiny použité k výpočtu společného spektra. Samozřejmě společně s nimi byl použit pro výpočet společného spektra i zkoumaný protein (lidský růstový hormon). Jak je vidět využívá se pouze receptor růstového hormonu z různých organismů. Na obrázku 50 je vidět signál zkoumaného proteinu, tedy lidského růstového hormonu. Na obrázku 51 je vidět společné spektrum proteinů z tabulky 14 a zpracovávaného proteinu (zobrazeného na obr. 50). Na něm je patrný jeden velmi výrazný vrchol a jeden zhruba třetinový. Hlavní vrchol se nachází na frekvenci 0,009259 Hz. Proteiny tedy mají jednu hlavní společnou funkci a jednu vedlejší společnou funkci. Na obrázku 52 je znázorněno amplitudové spektrum S-Transformace zkoumaného proteinu. A na obrázku 53 je

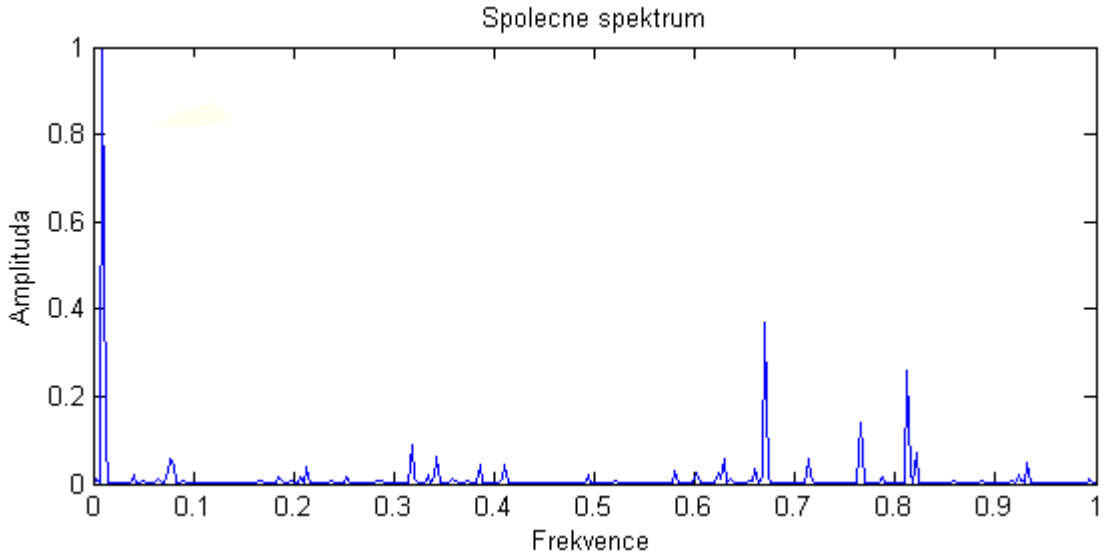
vidět amplitudové spektrum S-Transformace po vynásobení se společným spektrem (jeho plná velikost je v přílohách). Je opět zobrazena pouze pro délku zkoumané sekvence.

Tab. 14: Proteiny použité pro výpočet společného spektra

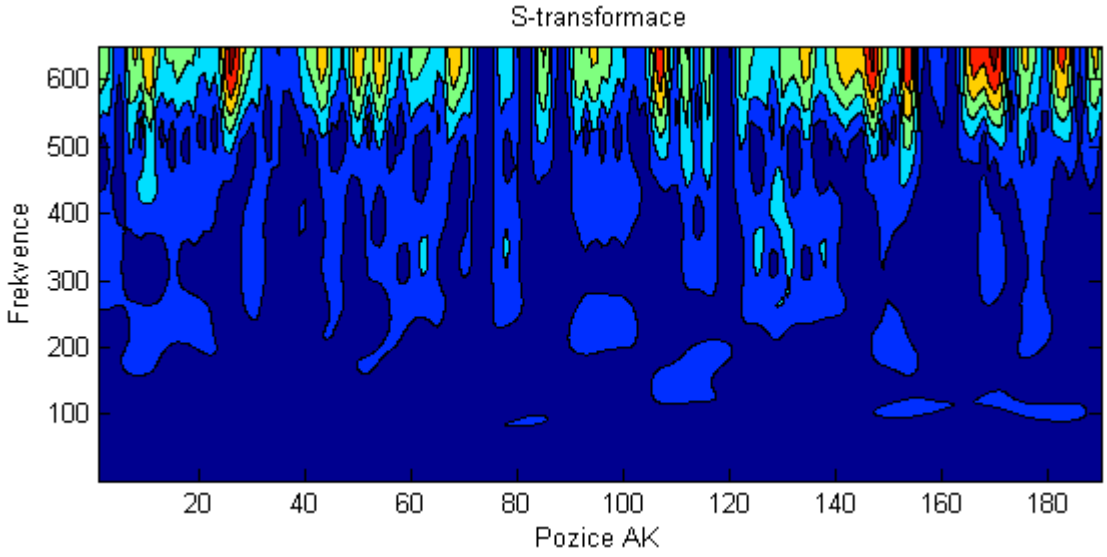
Název proteinu	**Organismus**	**Identif. Č.**
Receptor růstového hormonu	Homo sapiens	P10912
Receptor růstového hormonu	Rattus norvegicus	P16310
Receptor růstového hormonu	Mus musculus	P16882
Receptor růstového hormonu	Oryctolagus cuniculus	P19941
Receptor růstového hormonu	Cavia porcellus	Q9JI97
Receptor růstového hormonu	Canis familiaris	Q9TU69
Receptor růstového hormonu	Gallus gallus	Q02092
Receptor růstového hormonu	Bos taurus	O46600



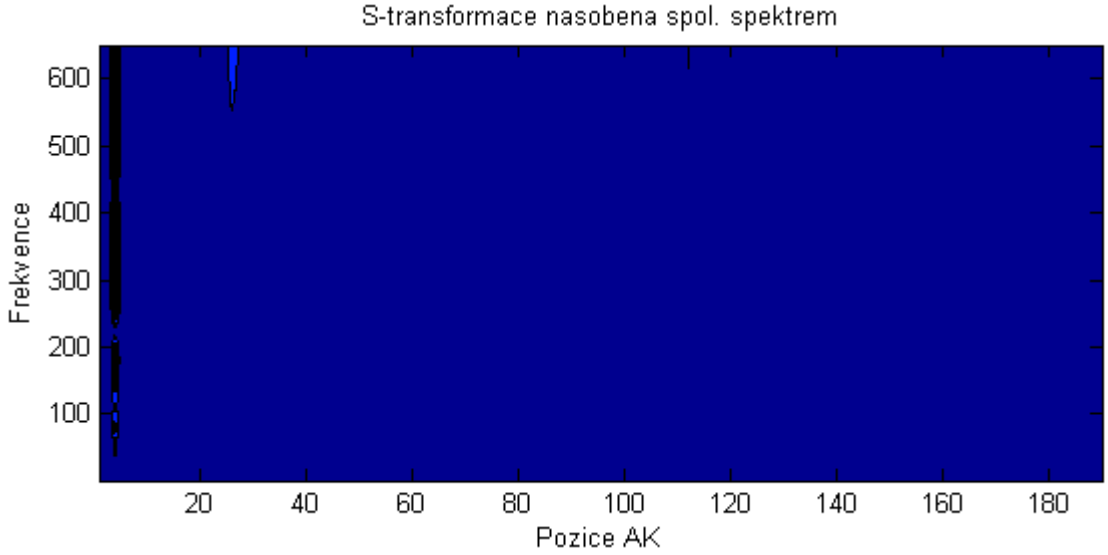
Obr. 50: Znázornění převedeného signálu (lidský růstový hormon) za využití hodnot EIIP



Obr. 51: Společné spektrum vypočítané z proteinů v tabulce 14 a lidského růstového hormonu

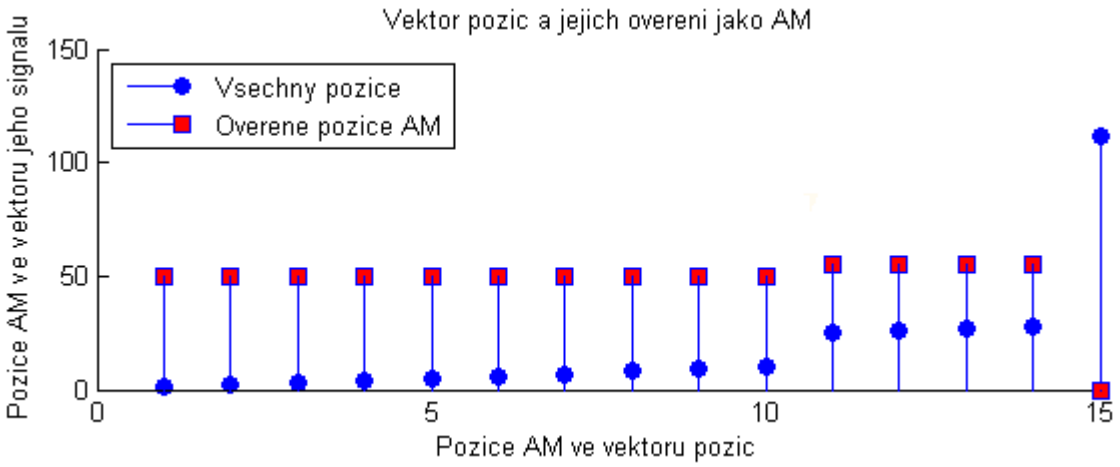


Obr. 52: Amplitudové spektrum S-Transformace zkoumaného signálu.

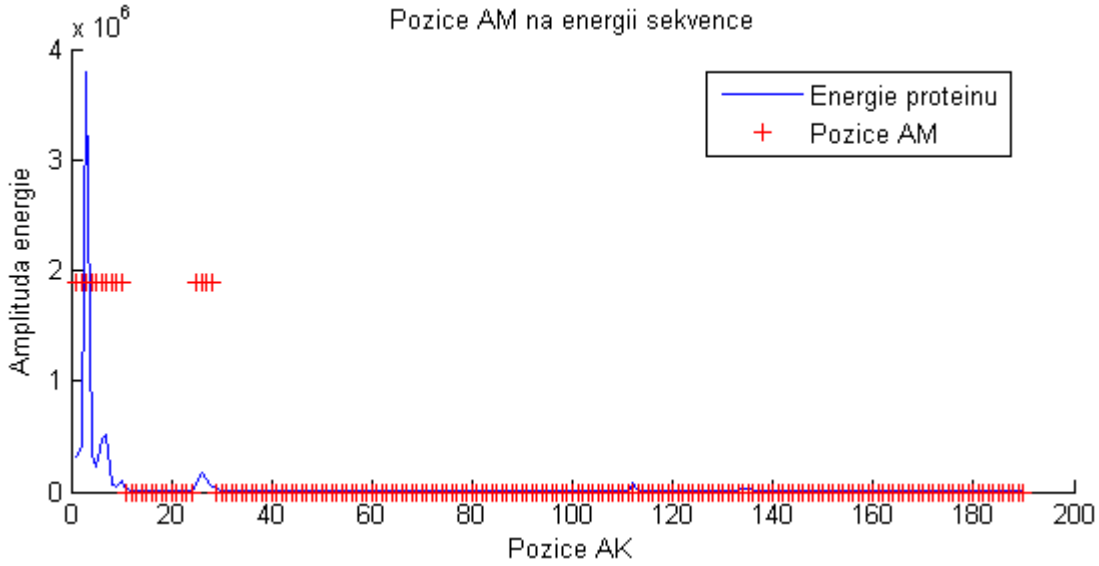


Obr. 53: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem

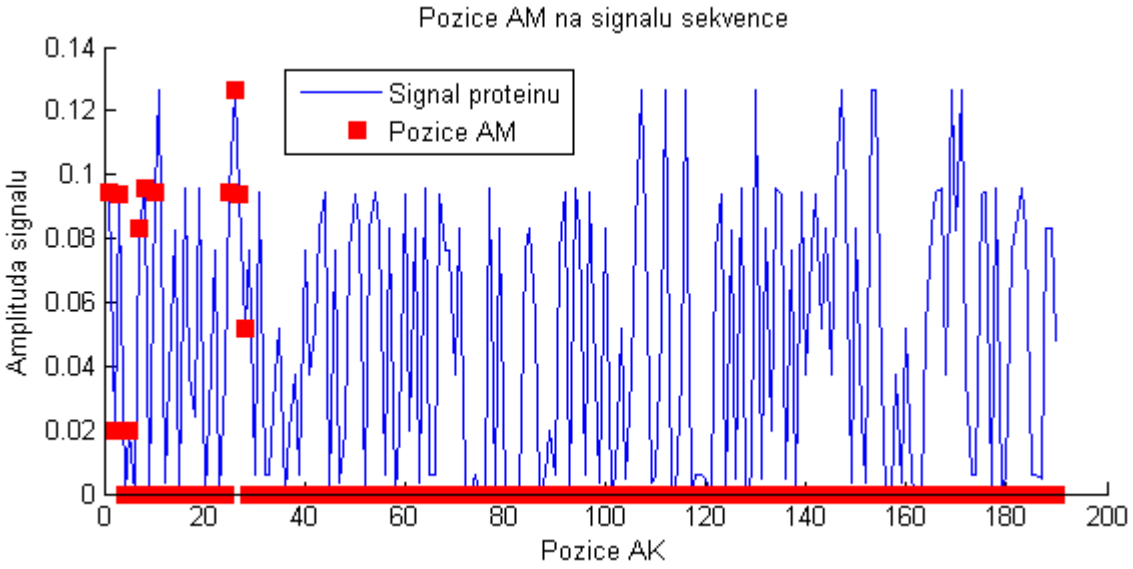
Na obrázku 54 je vidět vektor pozic a jejich ověření. Z něj vyplívá, že u tohoto zkoumaného signálu dostáváme dvě skupiny aktivních míst. První se nachází na pozicích 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 a ty ve fasta kódu sekvence představují znaky FPTIPLSRLF. Ty ukrývají aminokyseliny Phe-Pro-Thr-Ile-Pro-Leu-Ser-Arg-Leu-Phe. Druhá skupina pozic aktivních míst se nachází na 25, 26, 27, 28, ty ve fasta kódu sekvence představují znaky FDTY. Ty ukrývají následující aminokyseliny Phe-Asp-Thr-Tyr. Na obrázku 55 je vidět vektor energie signálu (po zpracování) a pozice aktivních míst. Na obrázku 56 je zobrazen původní signál a pozice ověřených aktivních míst.



Obr. 54: Vektory pozic a jejich ověření



Obr. 55: Vektor energie a pozice ověřených AM



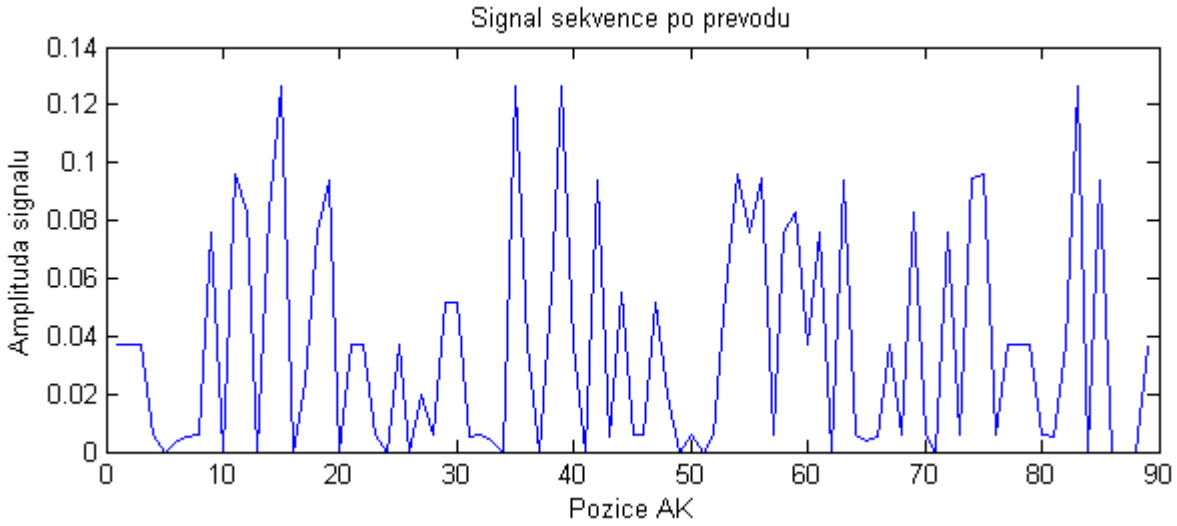
Obr. 56: Signál původní sekvence a pozice ověřených AM

6.6. Zpracování endonukleázy z Bacillus amyloliquefaciens (Barstar)

V tabulce 15 jsou uvedeny proteiny použité k výpočtu společného spektra. Samozřejmě společně s nimi byl použit pro výpočet společného spektra i zkoumaný protein (endonukleáza z Bacillus amyloliquefaciens). Jak je vidět využívá se podobných proteinů, či proteinů ze stejné rodiny, z různých organismů. Na obrázku 57 je vidět signál zkoumaného proteinu, tedy endonukleázy z Bacillus amyloliquefaciens. Na obrázku 58 je vidět společné spektrum proteinů z tabulky 15 a zpracovávaného proteinu (zobrazeného na obr. 57).

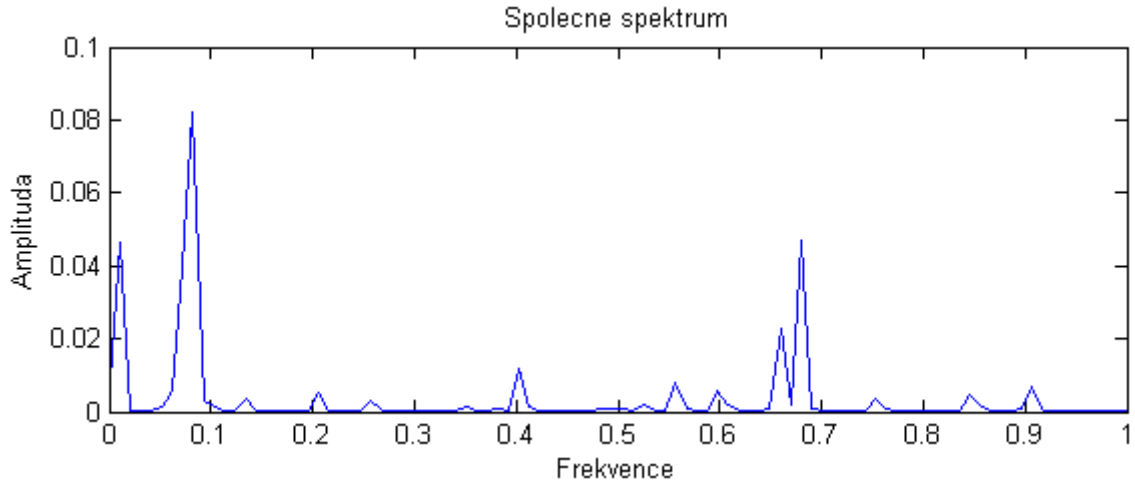
Tab. 15: Proteiny použité pro výpočet společného spektra

Název proteinu	**Organismus**	**Identif. Č.**
Inhibitor ribonukleázy	Bacillus amyloliquefaciens	P11540
Barstar	Yersinia pseudotuberculosis serotype	A7FDT9
Barstar	Yersinia pestis bv. Antiqua	A9R1V5
Barstar	Salmonella heidelberg	B4TJT7
Barstar	Salmonella enterica subsp.	B5PWB6
Barstar	Edwardsiella ictaluri	C5BAW5
Barnase inhibitor	Saccharomonospora viridis	C7MPS8
Protein z rodiny Barstar	Burkholderia thailandensis	Q2SZB1
Protein z rodiny Barstar	Burkholderia mallei	Q62H00

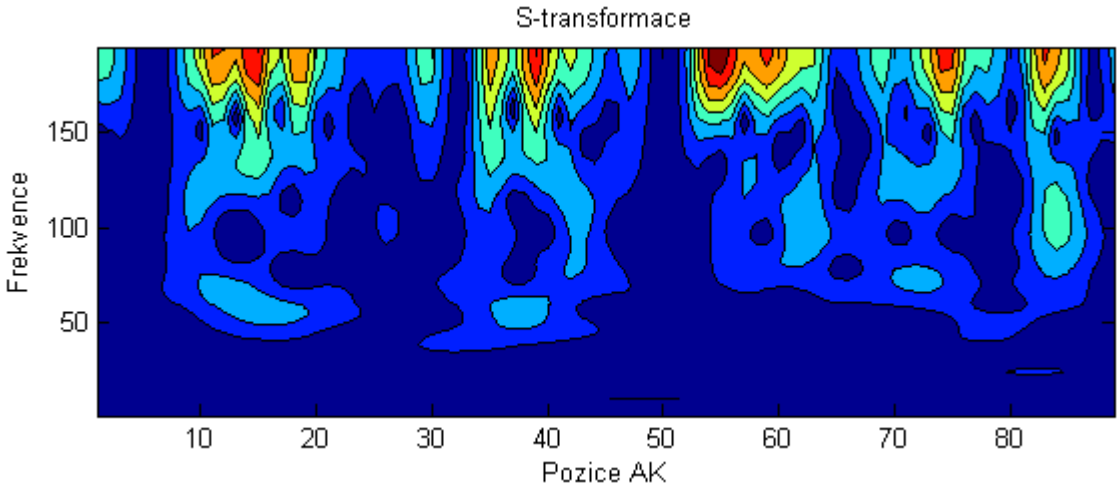


Obr. 57: Znázornění převedeného signálu (endonukleáza z Bacillus amyloliquefaciens - barstar) za využití hodnot EIIP

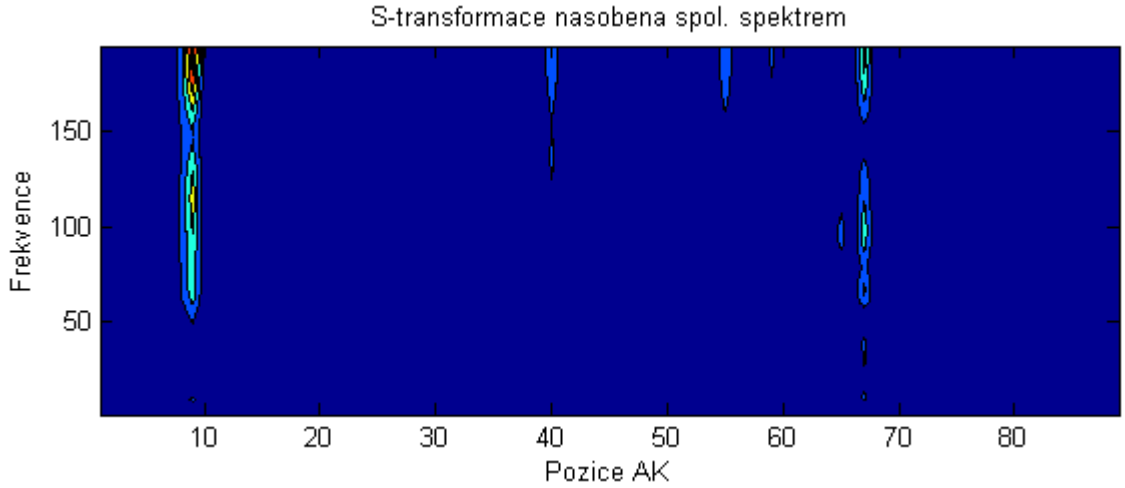
Jak je vidět na obrázku ukazující společné spektrum (obr. 58) je v něm jedem hlavní vrchol na frekvenci 0,08247 Hz a dva zhruba poloviční (na frekvencích 0,01031 Hz a 0,6804 Hz). Proteiny tedy mají sice společnou funkci ale asi mají více společných funkcí. Na obrázku 59 je vidět amplitudové spektrum S-Transformace zkoumaného signálu a na obrázku 60 je vidět amplitudové spektrum S-Transformace po vynásobení se společným spektrem (jeho plná velikost je v přílohách). Je opět zobrazena pouze pro délku zkoumané sekvence.



Obr. 58: Společné spektrum vypočítané z proteinů v tabulce 15 a endonukleázy z Bacillus amyloliquefaciens - barstar

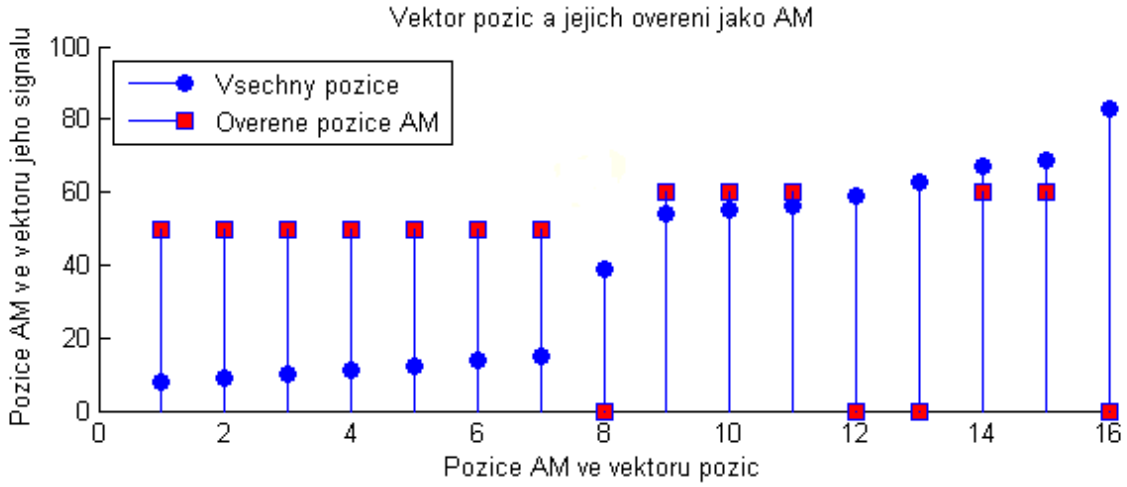


Obr. 59: Amplitudové spektrum S-Transformace zkoumaného signálu

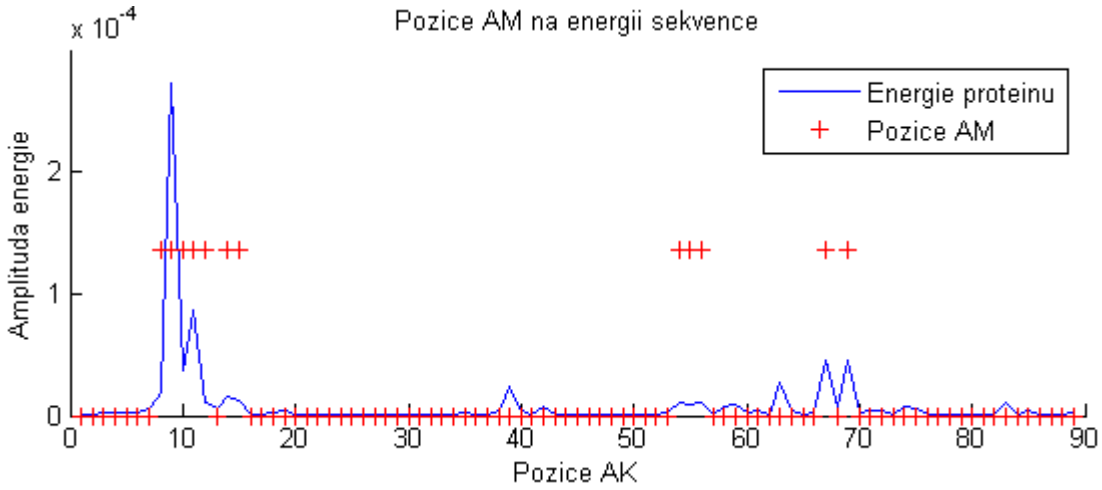


Obr. 60: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem

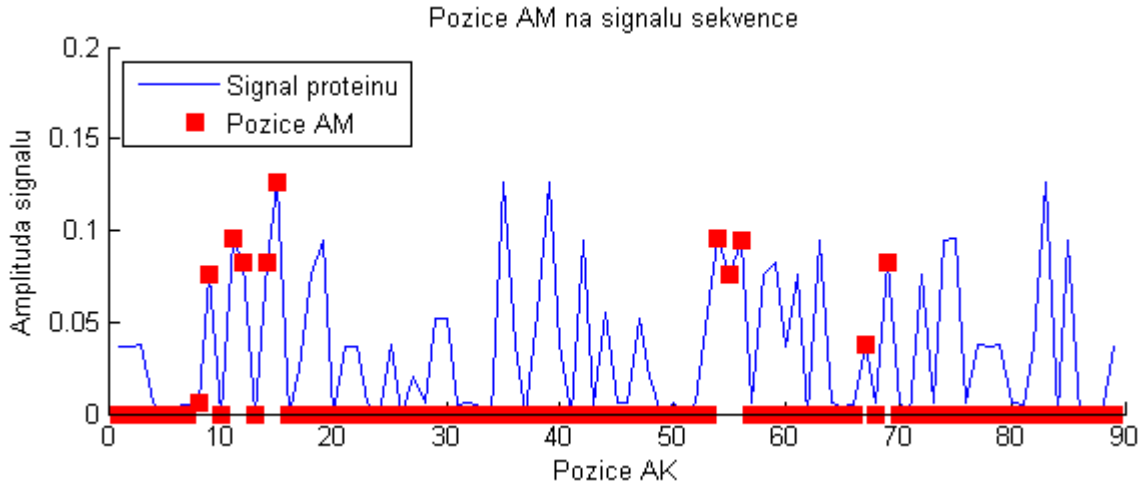
Na obrázku 61 je vidět vektor pozic a jejich ověření. Z něj vyplívá, že u tohoto zkoumaného signálu dostáváme tři skupiny aktivních míst. První se nachází na pozicích 8, 9, 10, 11, 12, 14, 15, a ty ve fasta kódu sekvence představují znaky EQIRS-SD. Ty ukrývají aminokyseliny Glu-Gln-Ile-Arg-Ser-GAP-Ser-Asp. Druhá skupina pozic aktivních míst se nachází na 54, 55, 56, ty ve fasta kódu sekvence představují znaky RQF. Ty ukrývají následující aminokyseliny Arg-Gln-Phe. Třetí skupina pozic se nachází na 67, 69, ty ve fasta kódu sekvence představují znaky A-S. Ty ukrývají následující aminokyseliny Ala-GAP-Ser. Na obrázku 62 je vidět vektor energie signálu (po zpracování) a pozice aktivních míst. Na obrázku 63 je zobrazen původní signál a pozice ověřených aktivních míst.



Obr. 61: Vektory pozic a jejich ověření



Obr. 62: Vektor energie a pozice ověřených AM



Obr. 63: Signál původní sekvence a pozice ověřených AM

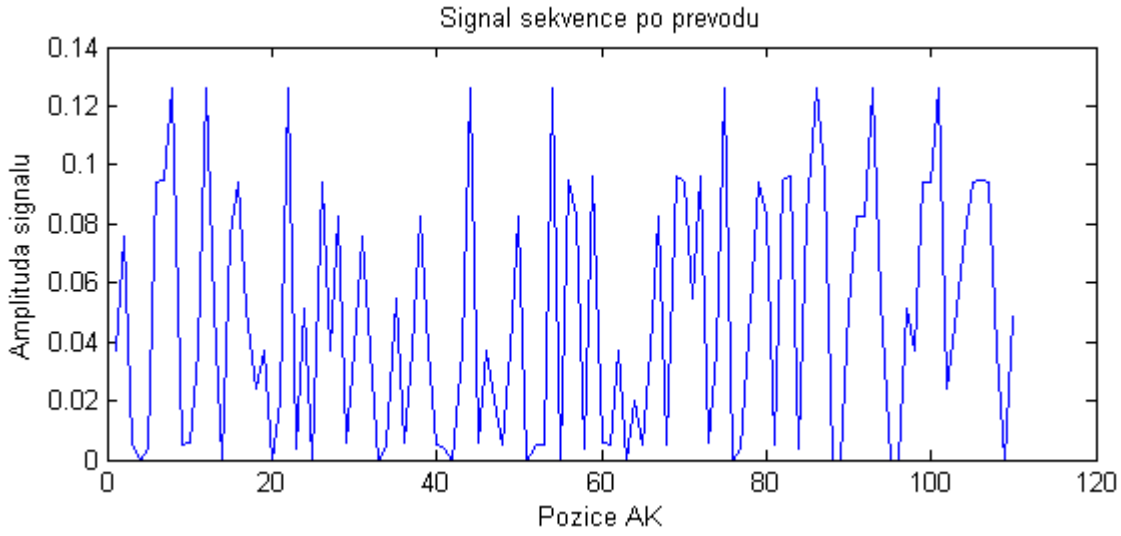
6.7. Zpracování endonukleázy z Bacillus amyloliquefaciens (Barnase)

V tabulce 16 jsou uvedeny proteiny použité k výpočtu společného spektra. Samozřejmě společně s nimi byl použit pro výpočet společného spektra i zkoumaný protein (endonukleáza z Bacillus amyloliquefaciens - barnase). Jak je vidět využívá se proteinů, či predikovaných proteinů. Nebo i proteinů údajných, tyto predikovanané a údajné se využívají hlavně u inhibitorů zkoumaného proteinu. Také se využívá různých organismů. Na obrázku 64 je vidět signál zkoumaného proteinu, tedy endonukleázy z Bacillus amyloliquefaciens. Na obrázku 65 je vidět společné spektrum proteinů z tabulky 16 a zpracovávaného proteinu (zobrazeného na obr. 64).

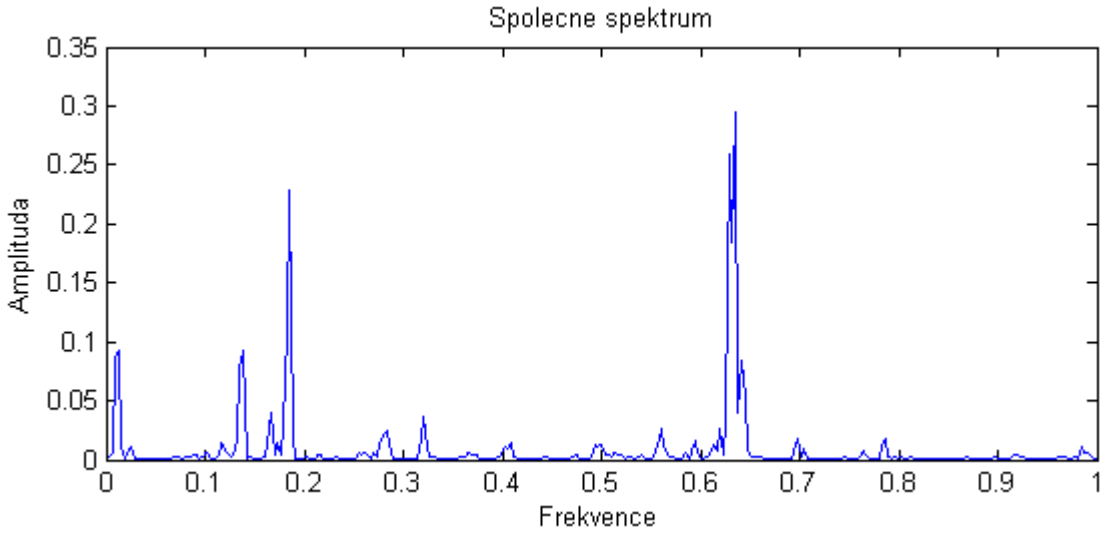
Tab. 16: Proteiny použité pro výpočet společného spektra

Název proteinu	**Organismus**	**Identif. Č.**
Receptor růstového hormonu	Homo sapiens	P10912
Ribonucleáza	Bacillus amyloliquefaciens	P00648
Barnase inhibitor	Pectobacterium wasabiae	D0KFB0
Barnase inhibitor	Streptomyces flavogriseus	C9NF27
Údajný barnase inhibitor	Hirschia baltica	C6XRM1
Pčedpoládaný barnase inhibitor	Escherichia coli	C6UF64
Barnase inhibitor	Thauera sp.	C4ZK78
Údajný barnase inhibitor	Escherichia coli	B7M0V1

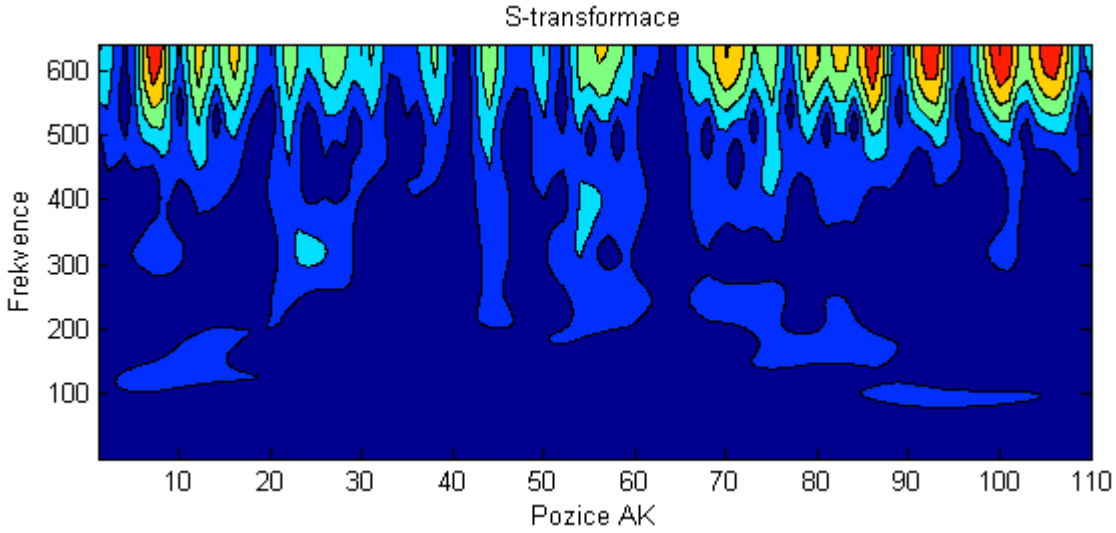
Jak je vidět na obrázku 65 ve společném spektru je více výrazných vrcholů. Dva nejvýraznější se nachází blízko u sebe. Ten největší se nachází na frekvenci 0,6352 Hz a u něj se nachází druhý největší vrchol na frekvenci 0,6289. Třetí nejvýraznější se nachází na frekvenci 0,1855 Hz. Tato přítomnost více vrcholů je nejspíše způsobena různorodostí proteinů zpracovávaných ve společném spektru.



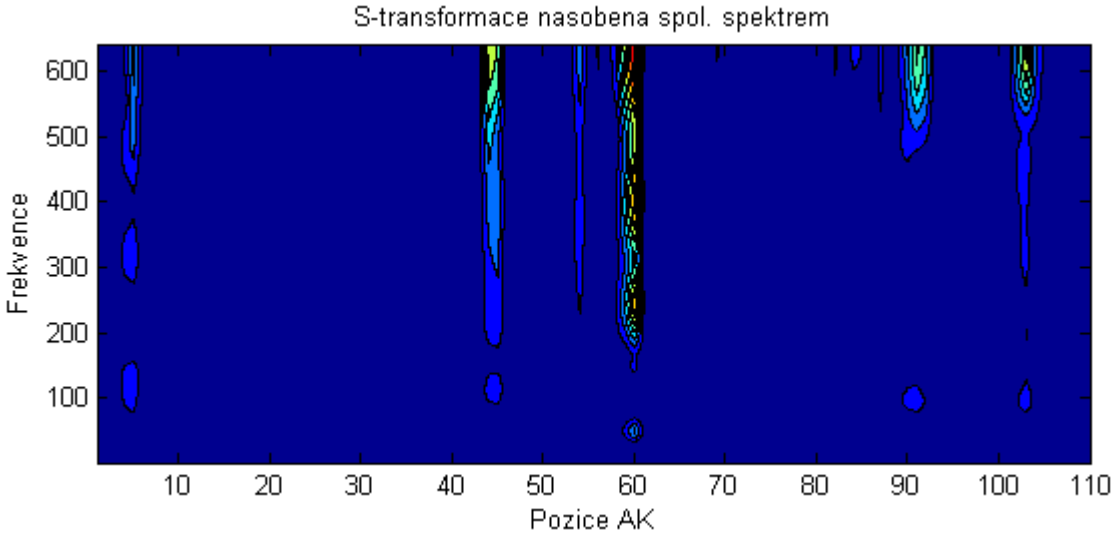
Obr. 64: Znázornění převedeného signálu (endonukleáza z Bacillus amyloliquefaciens - barnase) za využití hodnot EIIP



Obr. 65: Společné spektrum vypočítané z proteinů v tabulce 16 a endonukleázy z Bacillus amyloliquefaciens - barnase

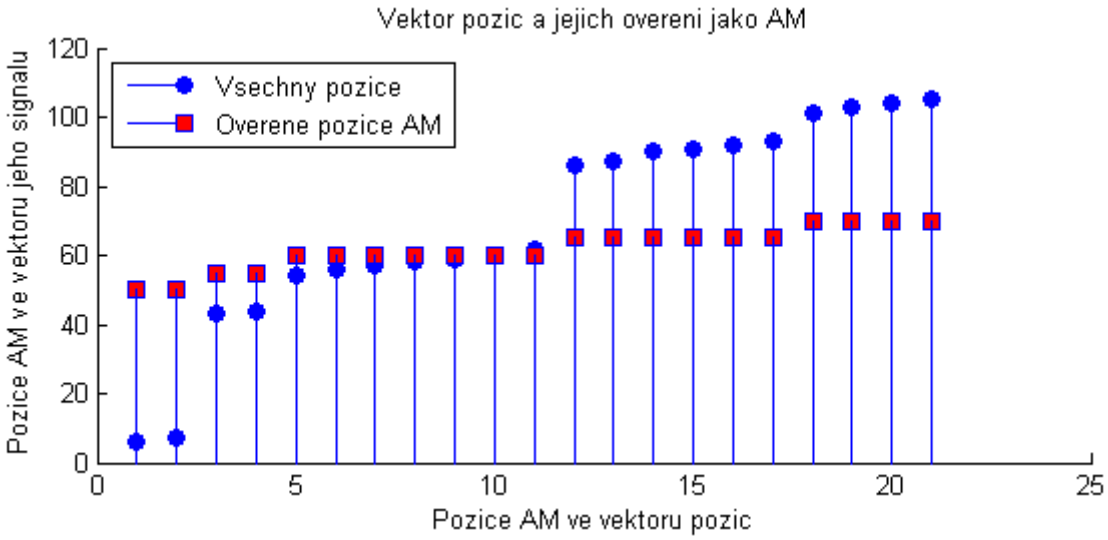


Obr. 66: Amplitudové spektrum S-Transformace zkoumaného signálu



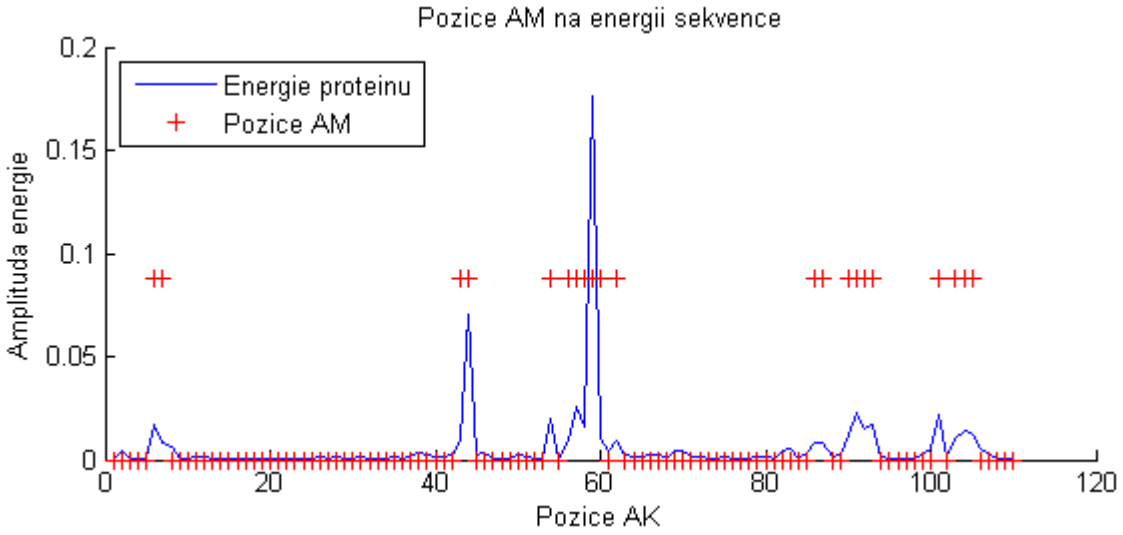
Obr. 67: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem

Na obrázku 66 je vidět amplitudové spektrum S-Transformace zkoumaného signálu a na obrázku 67 je vidět amplitudové spektrum S-Transformace po vynásobení se společným spektrem (jeho plná velikost je v přílohách). Je opět zobrazena pouze pro délku zkoumané sekvence. Na obrázku 68 je vidět vektor pozic a jejich ověření.

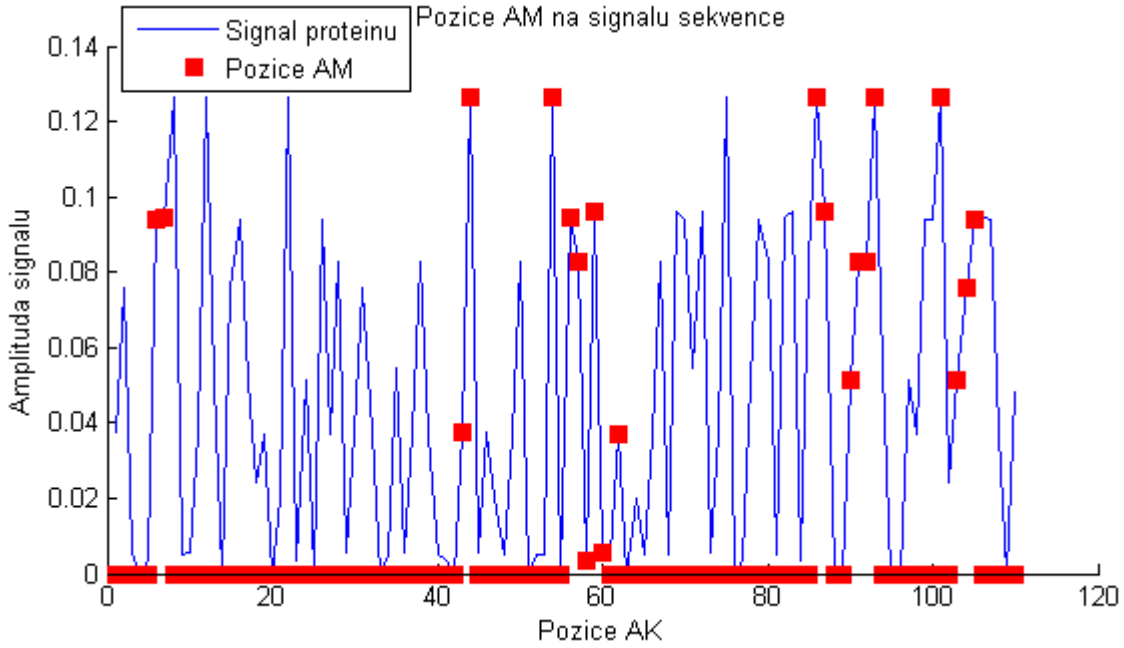


Obr. 68: Vektory pozic a jejich ověření

Z obrázku 68 vyplívá že u tohoto proteinu je identifikováno pět skupin aminokyselin identifikovaných jako aktivní místa. První skupina se nachází na pozicích 6,7 což ve fasta kódu sekvence představuje TF. Ty ukrývají aminokyseliny Thr-Phe. Druhá skupina je na pozicích 43, 44. Ty ve fasta kódu sekvence představují AD a ty reprezentují aminokyseliny Ala-Asp. Třetí skupina se rozkládá na pozicích 54, 56, 57, 58, 59, 60. Což ve fasta kódu sekvence znamená D-FSNRE ty reprezentují aminokyseliny Asp-GAP-Phe-Ser-Asn-Arg-Glu. Čtvrtá skupina je na pozicích 86, 87, 90, 91, 92, 93 ty ve fasta kódu sekvence představují znaky DR--YSSD. Ty reprezentují aminokyseliny Asp-Arg-GAP-GAP-Tyr-Ser-Ser-Asp. Poslední, pátá, skupina se nachází na pozicích 101, 103, 104, 105. Ty ve fasta kódu představují znaky D-YQT, ty ukrývají aminokyseliny Asp-GAP-Tyr-Gln-Thr. Na obrázku 69 je vidět je vidět vektor energie signálu (po zpracování) a pozice aktivních míst. Na obrázku 70 je zobrazen původní signál a pozice ověřených aktivních míst.



Obr. 69: Vektor energie a pozice ověřených AM



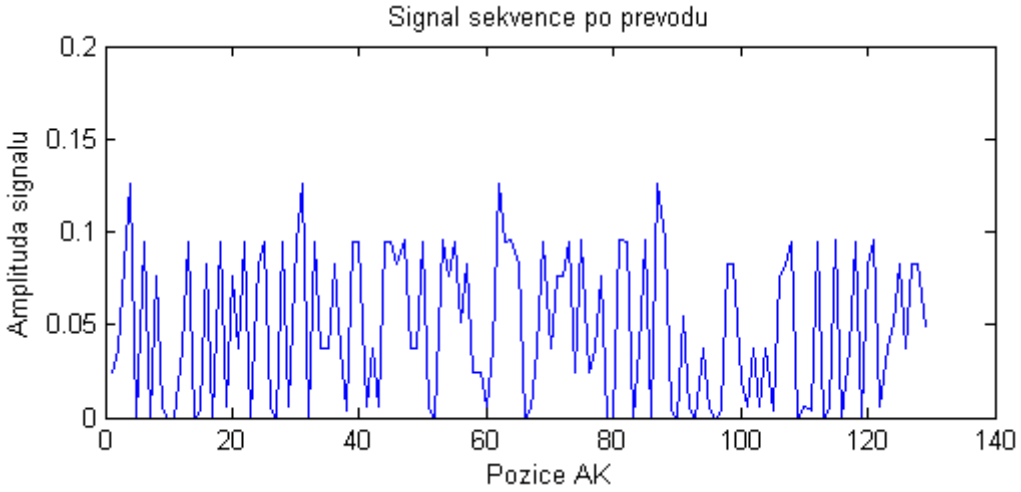
Obr. 70: Signál původní sekvence a pozice ověřených AM

6.8. Zpracování lidského proteinu Interleukin - 4

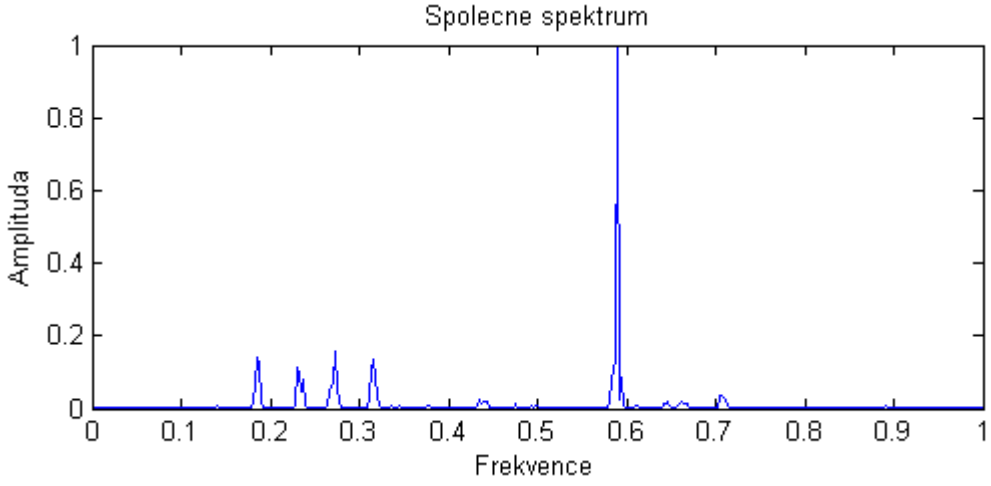
V tabulce 17 jsou uvedeny proteiny použité k výpočtu společného spektra. Samozřejmě společně s nimi byl použit pro výpočet společného spektra i zkoumaný protein (lidský Interleukin - 4). Jak je vidět jsou využity stejné proteiny, ale od různých organismů. Na obrázku 71 je vidět signál zkoumaného proteinu, tedy lidský Interlukin - 4. Na obrázku 72 je vidět společné spektrum proteinů z tabulky 16 a zpracovávaného proteinu (zobrazeného na obr. 71).

Tab. 17: Proteiny použité pro výpočet společného spektra

Název proteinu	**Organismus**	**Identif. Č.**
Interleukin-4	Homo sapiens	P05112
Interleukin-4	Mus musculus	P07750
Interleukin-4 receptor subunit sloha	Mus musculus	P16382
Interleukin-4	Rattus norvegicus	P20096
Interleukin-4	Ovis aries	P30368
Interleukin-4	Equus caballus	P42202
Interleukin-4	Cervus elaphus	P51744
Interleukin-4	Pan troglodytes	Q8HYB1
Interleukin-4	Sus scrofa	Q04745



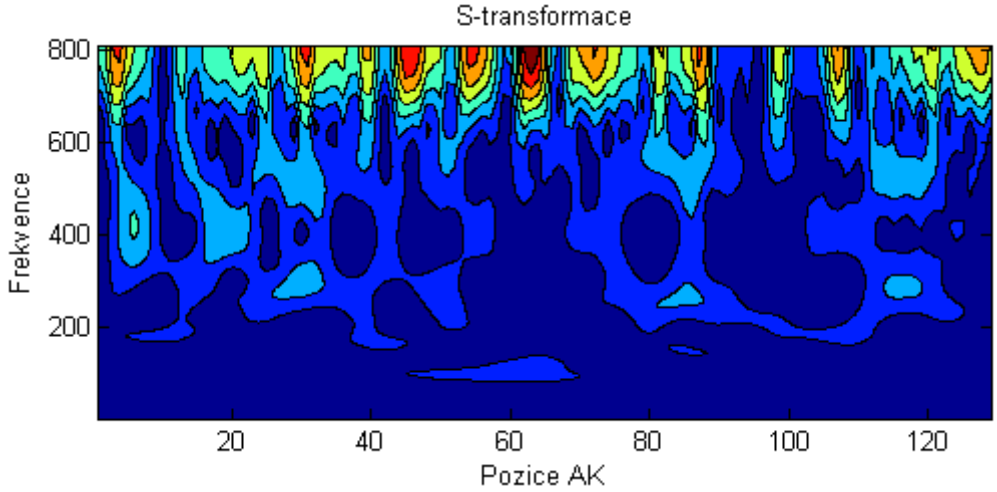
Obr. 71: Znázornění převedeného signálu (lidský Interleukin – 4) za využití hodnot EIIP



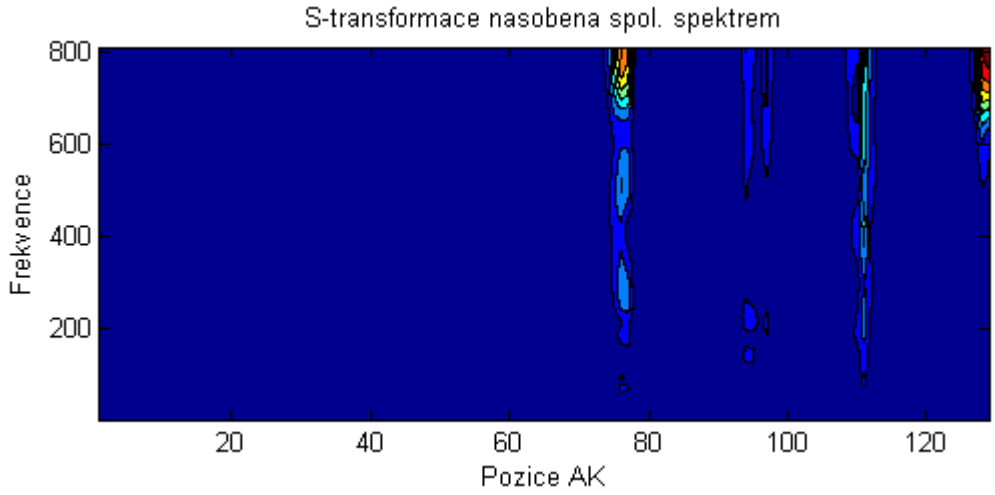
Obr. 72: Společné spektrum vypočítané z proteinů v tabulce 17 a lidského Interleukin - 4

Jak je vidět na obrázku 72 je ve společném spektru je pouze jeden výrazný vrchol. Ten se nachází na frekvenci 0,5891 Hz. Na obrázku 73 je vidět amplitudové spektrum S-Transformace zkoumaného signálu a na obrázku 74 je vidět amplitudové spektrum S-Transformace po vynásobení se společným spektrem (jeho plná velikost je v přílohách). Je

opět zobrazena pouze pro délku zkoumané sekvence. Na obrázku 75 je vidět vektor pozic a jejich ověření.

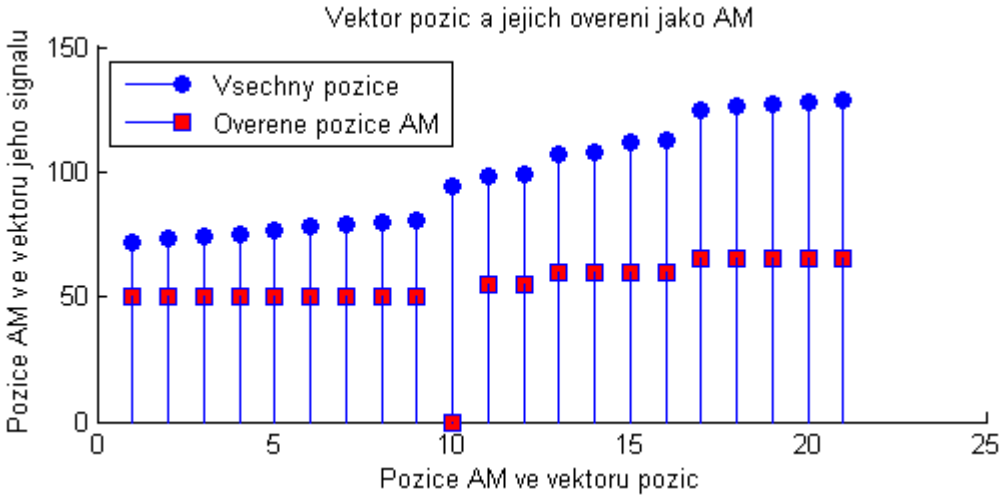


Obr. 73: Amplitudové spektrum S-Transformace zkoumaného signálu



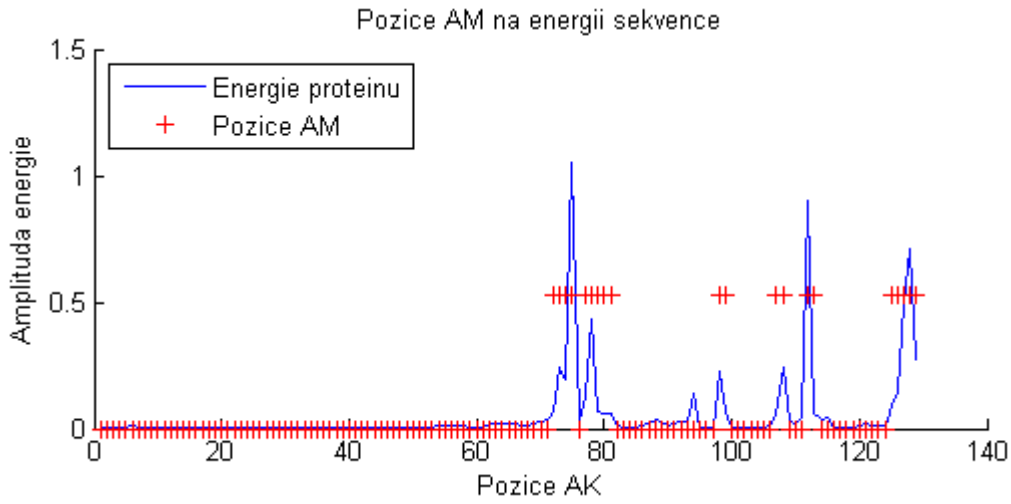
Obr. 74: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem

Jak je vidět na obrázku 75 je vidět, že je identifikováno pět skupin aminokyselin jako aktivní místa. První se nachází na pozicích 72, 73, 74, 75, 77, 78, 79, 80, 81. Ty ve fasta kódu sekvence znamenají znaky QFHR-KQLIR což reprezentuje aminokyseliny Gln-Phe-His-Arg-GAP-Lys-Gln-Leu-Ile-Arg. Druhá skupina se nachází na pozicích 98, 99 což ve fasta kódu sekvence znamená SC. To ukrývá aminokyseliny Ser-Cys. Třetí skupina je na pozicích 107, 108 ty ve fasta kódu sekvence představují ST. Což reprezentuje aminokyseliny Ser-Thr. Čtvrtá skupina je pozicích 112, 113 a ve fasta kódu sekvence se na těchto pozicích nachází znaky FL. To představuje aminokyseliny Phe-Leu. A poslední pátá skupina se nachází na pozicích 125, 126, 127, 128, 129. Ve fasta kódu sekvence tedy znaky SKCSS. Což reprezentuje aminokyseliny Ser-Lys-Cys-Ser-Ser.

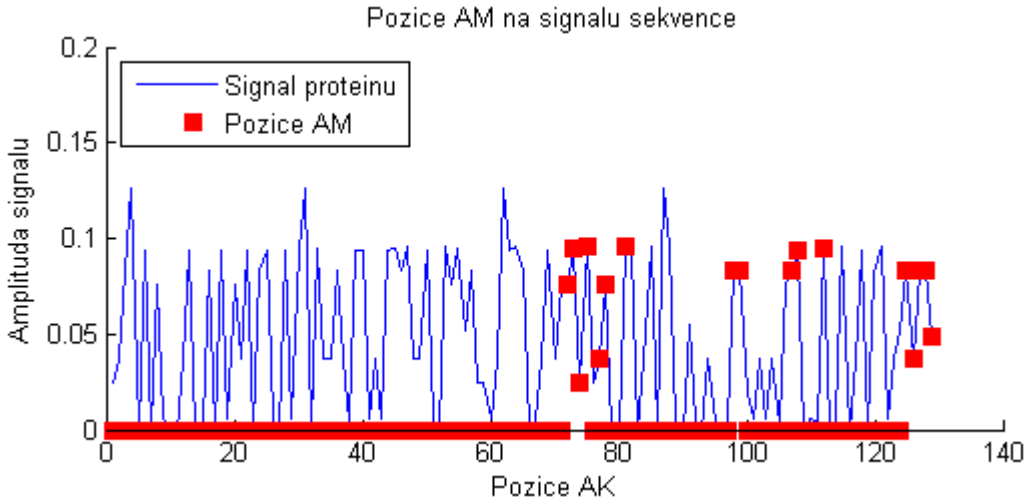


Obr. 75: Vektory pozic a jejich ověření

Na obrázku 76 je vidět je vidět vektor energie signálu (po zpracování) a pozice aktivních míst. Na obrázku 77 je zobrazen původní signál a pozice ověřených aktivních míst.



Obr. 76: Vektor energie a pozice ověřených AM



Obr. 77: Signál původní sekvence a pozice ověřených AM

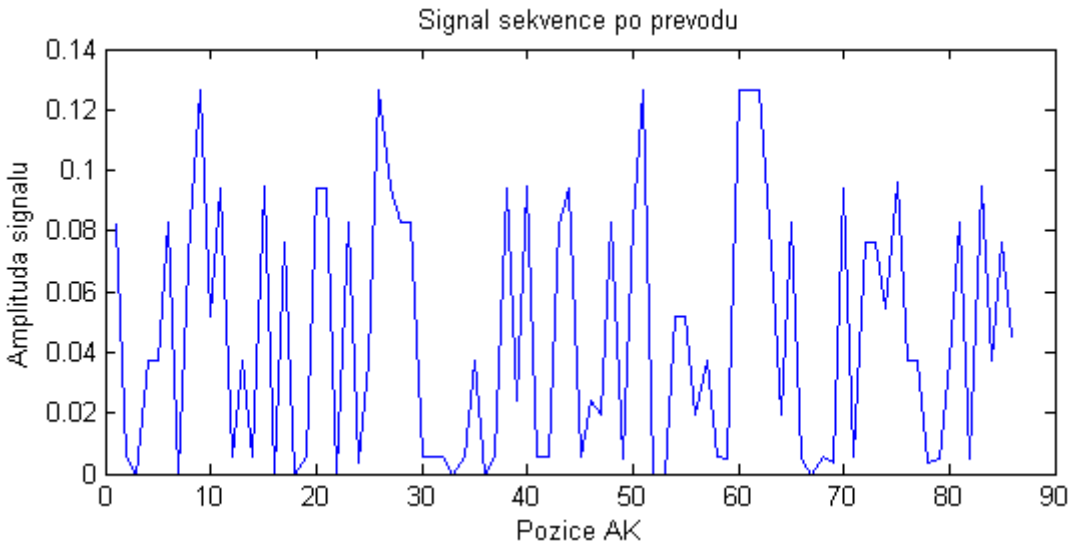
6.9. Zpracování colicinu E9 imunitního proteinu z Escherichia coli

V tabulce 18 jsou uvedeny proteiny použité k výpočtu společného spektra. Samozřejmě společně s nimi byl použit pro výpočet společného spektra byl zkoumaný protein (Colicinu E9 z Escherichia coli).

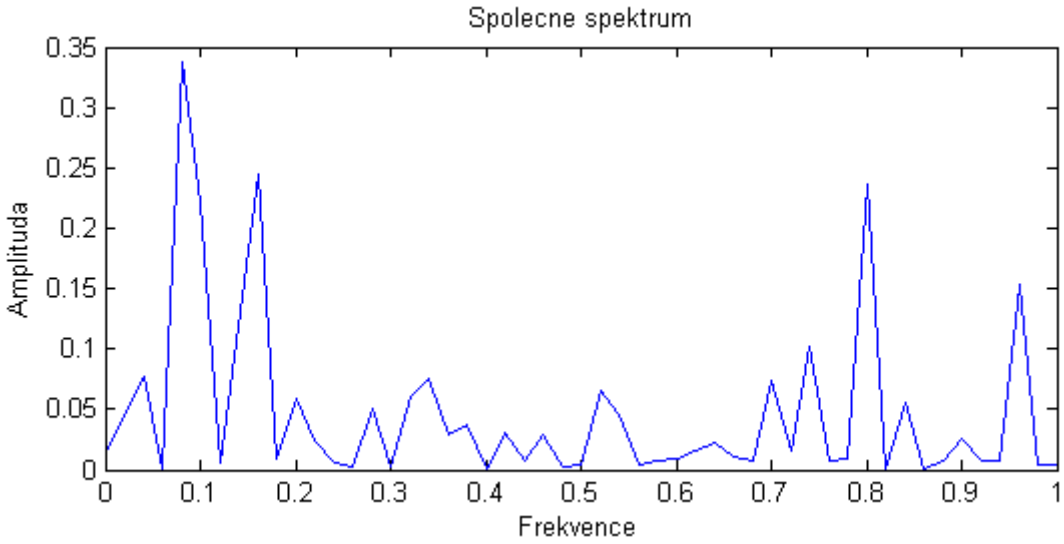
Tab. 18: Proteiny použité pro výpočet společného spektra

Název proteinu	**Organismus**	**Identif. Č.**
Colicin-E9 immunitní protein	Escherichia coli	P13479
Lysinový protein pro colicin E9	Escherichia coli	P15176
Předpokládaný protein	Populus trichocarpa	B9MVA0

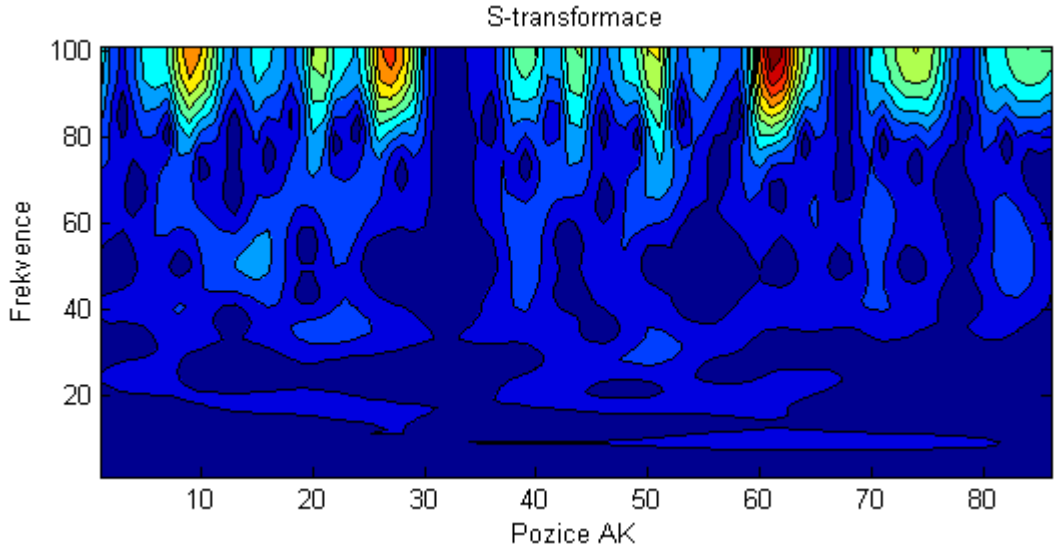
Jak je vidět z tabulky 18 využívá se poměrně málo proteinů. Navíc se využívají i proteiny předpokládané. Zkoumaný protein je vidět na obrázku 78. Společné spektrum je pak vidět na obrázku 79, jak je na něm jeden hlavní vrchol na frekvenci 0,08 Hz a dva menší vrcholy na frekvencích 0,16 Hz a 0,8 Hz. Na obrázku 80 je vidět amplitudové spektrum S-Transformace zkoumaného signálu a na obrázku 81 je vidět amplitudové spektrum S-Transformace po vynásobení se společným spektrem (jeho plná velikost je v přílohách). Je opět zobrazena pouze pro délku zkoumané sekvence. Na obrázku 82 je vidět vektor pozic a jejich ověření.



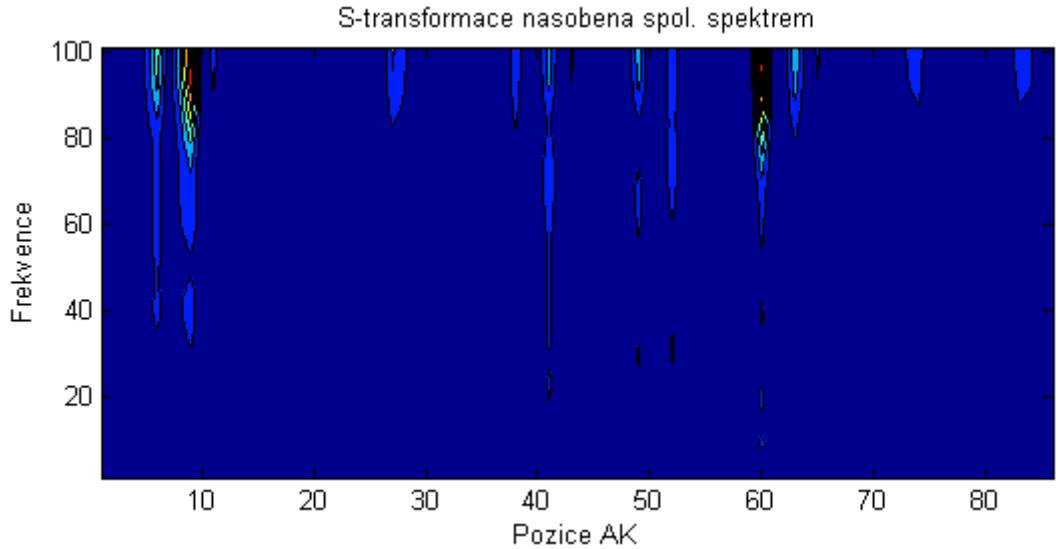
Obr. 78: Znázornění převedeného signálu (Colicinu E9 z Escherichia coli) za využití hodnot EIIP



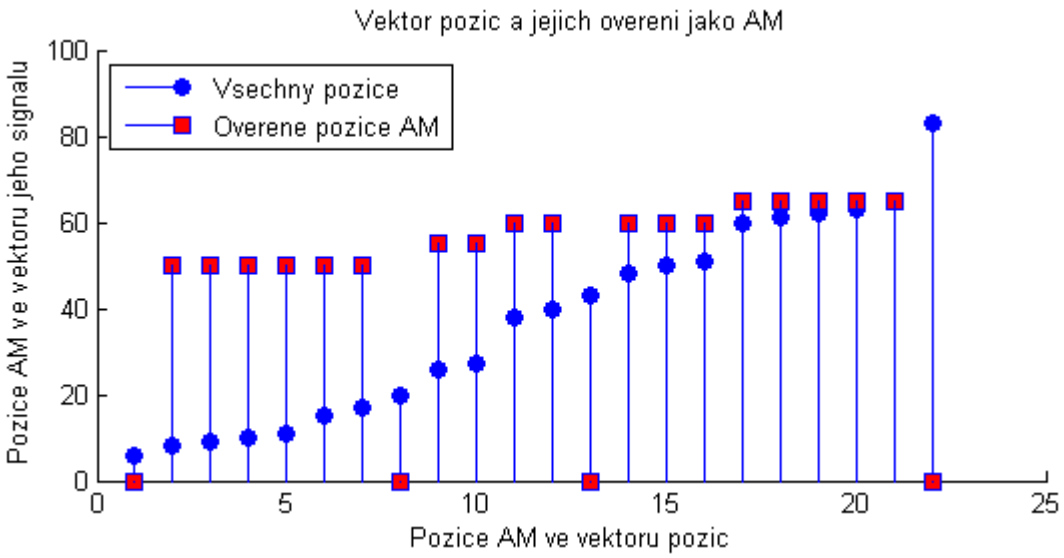
Obr. 79: Společné spektrum vypočítané z proteinů v tabulce 18 a Colicinu E9 z Escherichia coli



Obr. 80: Amplitudové spektrum S-Transformace zkoumaného signálu

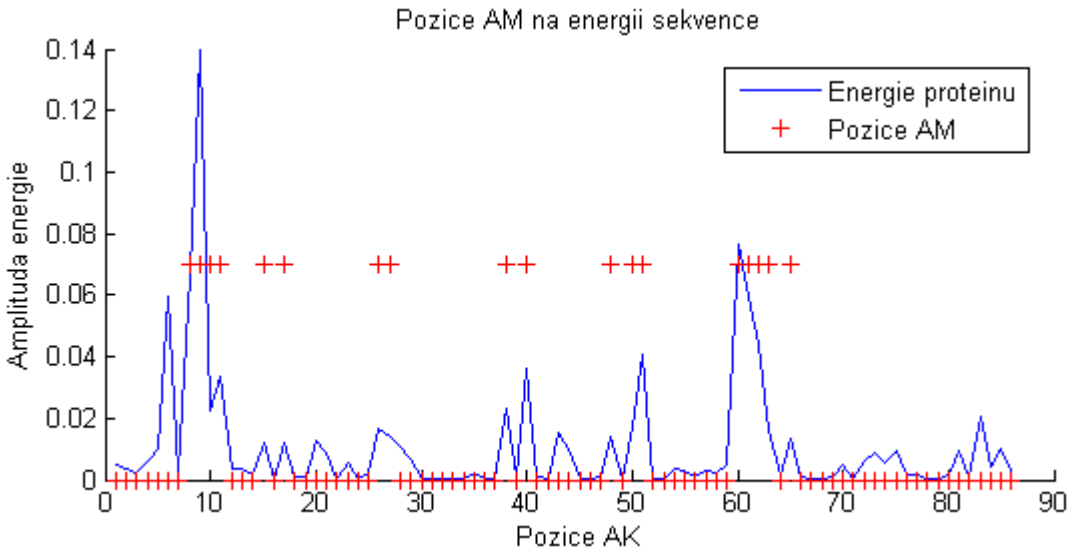


Obr. 81: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem

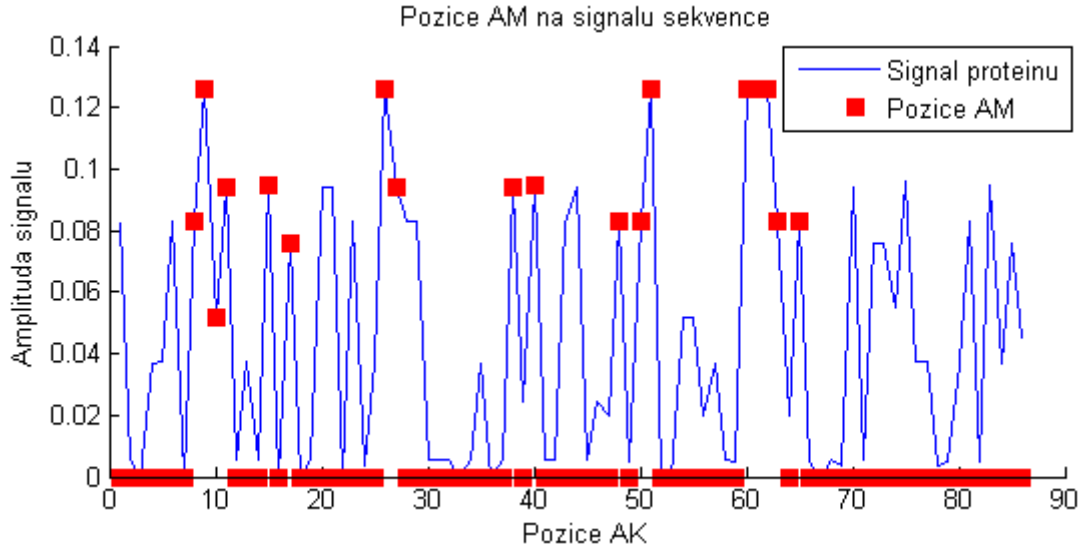


Obr. 82: Vektory pozic a jejich ověření

Jak je vidět na obrázku 82, je identifikováno šest skupin aminokyselin identifikovaných jako aktivní místa. První se nachází na pozicích 8, 9, 10, 11 ty ve fasta kódu sekvence představují znaky SDYT. Což představuje aminokyseliny Ser-Asp-Tyr-Thr. Druhá skupina se nachází na pozicích 15, 17 což ve fasta kódu sekvence znamená F-Q. To ukrývá aminokyseliny Phe-GAP-Gln. Třetí skupina je na pozicích 26, 27 ty ve fasta kódu sekvence představují DT. Což reprezentuje aminokyseliny Asp-Thr. Čtvrtá skupina je pozicích 38, 40 a ve fasta kódu sekvence se na těchto pozicích nachází znaky T-F. To představuje aminokyseliny Thr-GAP-Phe. Pátá skupina se nachází na pozicích 48, 50, 51. Ve fasta kódu sekvence tedy znaky S-SD. Což reprezentuje aminokyseliny Ser-GAP-Ser-Asp. A poslední šestá skupina je na pozicích 60, 61, 62, 63, 65. Ve fasta kódu sekvence tedy DDDS-S což představuje aminokyseliny Asp-Asp-Asp-Ser-GAP-Ser. Na obrázku 83 je vidět je vidět vektor energie signálu (po zpracování) a pozice aktivních míst. Na obrázku 84 je zobrazen původní signál a pozice ověřených aktivních míst.



Obr. 83: Vektor energie a pozice ověřených AM



Obr. 84: Signál původní sekvence a pozice ověřených AM

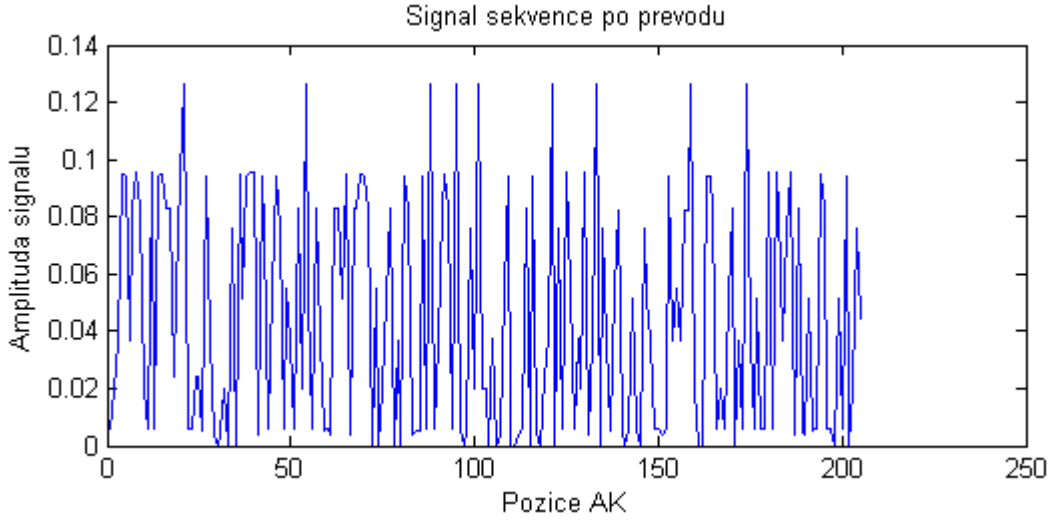
6.10. Zpracování receptor lidského růstového hormonu

V tabulce 19 jsou uvedeny proteiny použité k výpočtu společného spektra. Samozřejmě společně s nimi byl použit pro výpočet společného spektra i zkoumaný protein (receptor lidského růstového hormonu).

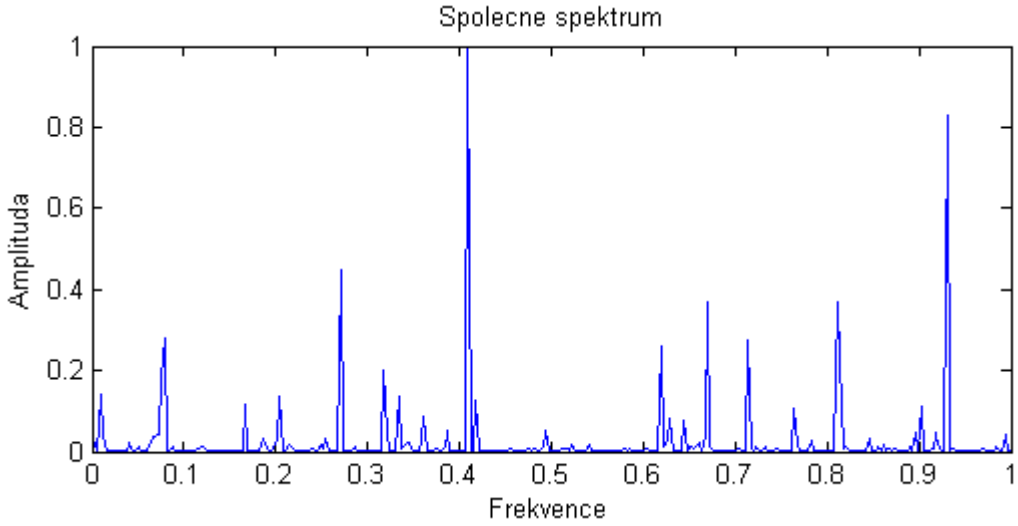
Tab. 19: Proteiny použité pro výpočet společného spektra

Název proteinu	**Organismus**	**Identif. Č.**
Receptor růstového hormonu	Bos indicie	P79108
Receptor růstového hormonu	Macaca mulatka	P79194
Receptor růstového hormonu	Papio anubis	Q9XSZ1
Receptor růstového hormonu	Ailuropoda melanoleuca	Q95JF2
Receptor růstového hormonu	Saimiri boliviensis boliviensis	Q95ML5
Receptor růstového hormonu	Ovis aries	Q28575

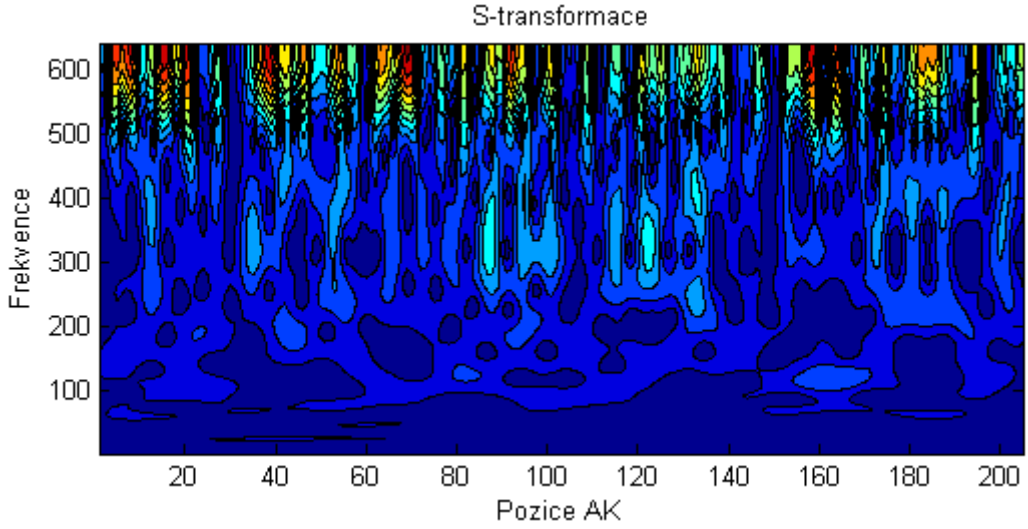
Jak je vidět z tabulky 19 využívá se stejný protein z různých organismů. Zkoumaný protein je vidět na obrázku 85. Společné spektrum je pak vidět na obrázku 86, jak je na něm jeden hlavní vrchol na frekvenci 0,4088 Hz a jeden o něco menší vrchol na frekvenci 0,9308 Hz. Na obrázku 87 je vidět amplitudové spektrum S-Transformace zkoumaného signálu a na obrázku 88 je vidět amplitudové spektrum S-Transformace po vynásobení se společným spektrem (jeho plná velikost je v přílohách). Je opět zobrazena pouze pro délku zkoumané sekvence. Na obrázku 89 je vidět vektor pozic a jejich ověření.



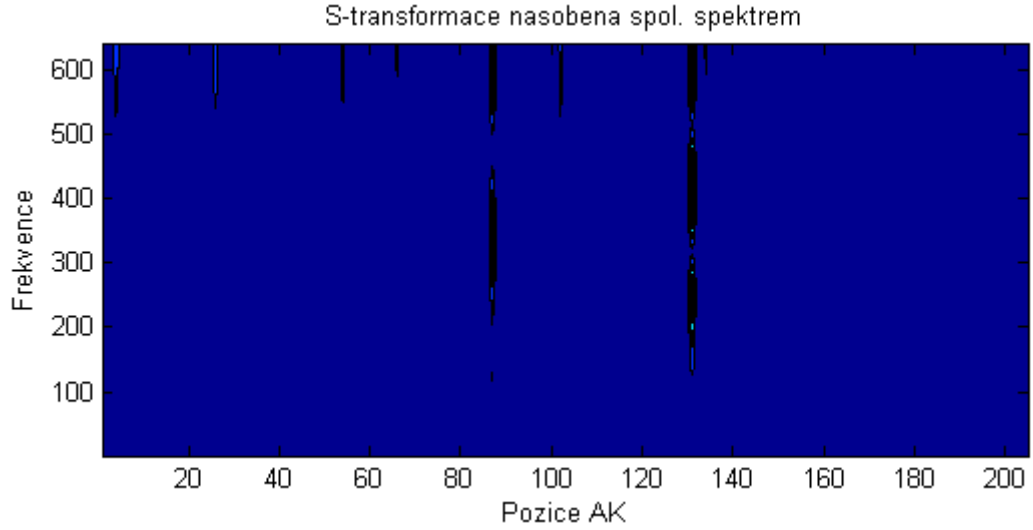
Obr. 85: Znázornění převedeného signálu (receptor lidského růstového hormonu) za využití hodnot EIIP



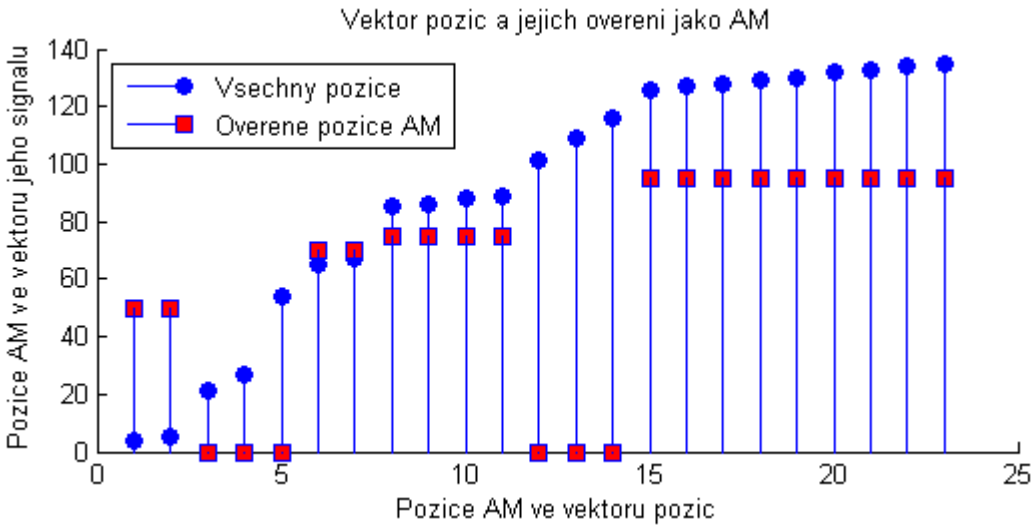
Obr. 86: Společné spektrum vypočítané z proteinů v tabulce 19 a receptoru lidského růstového hormonu



Obr. 87: Amplitudové spektrum S-Transformace zkoumaného signálu

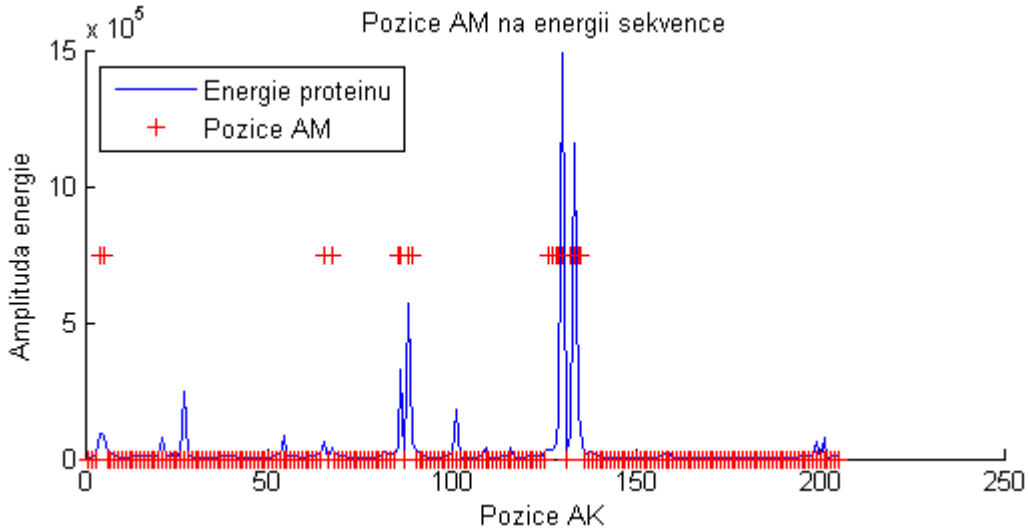


Obr. 88: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem

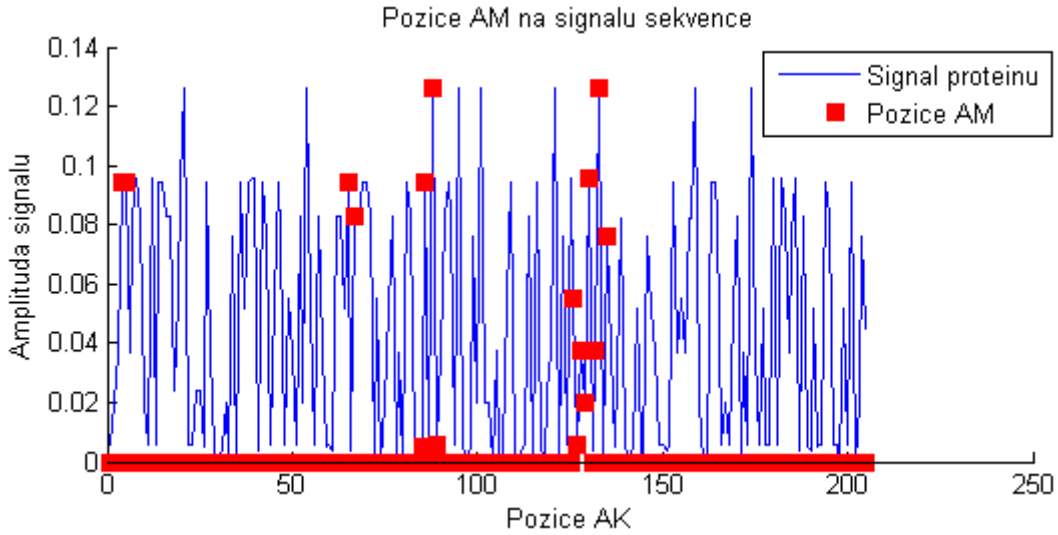


Obr. 89: Vektory pozic a jejich ověření

Jak je vidět na obrázku 89, jsou identifikovány čtyři skupiny aminokyselin identifikovaných jako aktivní místa. První se nachází na pozicích 4,5 ty ve fasta kódu sekvence představují znaky FT. Což představuje aminokyseliny Phe-Thr. Druhá skupina se nachází na pozicích 65, 67 což ve fasta kódu sekvence znamená F-S. To ukrývá aminokyseliny Phe-GAP-Ser. Třetí skupina je na pozicích 85, 86, 88, 89 ty ve fasta kódu sekvence představují GT-DE. Což reprezentuje aminokyseliny Gly-Thr-GAP-Asp-Glu. A poslední čtvrtá skupina je na pozicích 126, 127, 128, 129, 130, 132, 133, 134, 135. Ve fasta kódu sekvence tedy WEAPR-ADIQ což představuje aminokyseliny Trp-Glu-Ala-Pro-Arg-GAP-Ala-Asp-Ile-Gln. Na obrázku 90 je vidět je vidět vektor energie signálu (po zpracování) a pozice aktivních míst. Na obrázku 91 je zobrazen původní signál a pozice ověřených aktivních míst.



Obr. 90: Vektor energie a pozice ověřených AM



Obr. 91: Signál původní sekvence a pozice ověřených AM

6.11. Souhrn výsledků

V tabulce 20 jsou uvedeny všechna pozice zkoumaných proteinů označených jako aktivní místa. V tabulce 21 je vidět porovnání tří metod (Robetta-Ala, S-transformace z [3] a mnou provedené S-Trasnformace) a databáze (ASEdb), pro pět proteinů (lidský růstový hormon, receptor lidského růstového hormonu, Endonukleázy z bacillus amyloliquefaciens – barstar i Barnase a Colicin E9 imunitní protein z Escherichia coli). A v tabulce 22 je vidět výkon mé S-Transformace na jednotlivých zkoumaných proteinech. Kde se jako referenční zdroj aktivních míst použije ASEdb spolu s Robetta-Ala. Dále v tabulce 23 je vidět porovnání celkového výkonu mé S-Transformace a S-Transformace z [3]. V tabulce 24 jsou vidět výkony mé metody, pokud se jako reference vezme původní meotda z [3]. Na konec je v tabulce 25 ukázáno porovnání s dostupnými metodami. Kde jako referenční data byla brána metoda Robetta – Ala a databáze ASEdb.

Tab. 20: Pozice aktivních míst

Název Proteinu	**Pozice aktivních míst**
Fibroblásotvý růstový faktor	9, 10, 11, 12, 13 ,14, 15, 16, 17, 87, 89, 90, 120,121,128, 130
Endonukleáza C	21, 22, 26 ,27, 40, 42, 43, 58, 59, 60, 61, 70, 71, 73, 75, 96, 97, 100, 101, 103, 106, 107, 108, 109, 111,119, 120, 132, 133, 134, 135, 137, 138, 140, 142, 143
TRP RNA-Vazebný útlomový protein	1, 2, 3, 4, 5, 6, 7, 8, 71, 72, 73, 74, 75
Lidský alpha hemoglobin	1, 2, 3, 4, 5, 6, 7, 31, 32, 33, 138, 139, 140, 141
Lidský růstový hormon	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 25, 26, 27, 28
Endonukleáza (Barstar)	8, 9, 10, 11, 12, 14, 15, 54, 55, 56, 67, 69
Endonukleáza (Barnase)	6, 7, 43, 44, 54, 56, 57, 58, 59, 60, 86, 87, 90, 91, 92, 93, 101, 103, 104, 105
Interleukin – 4	72, 73, 74, 75, 77, 78, 79, 80, 81, 98, 99, 107, 108, 112, 113, 125, 126, 127, 128, 129
Colicin E9 imunitní protein	8, 9, 10, 11, 15, 17, 26, 27, 38, 40, 48, 50, 51, 60, 61, 62, 63, 65
Lidský růstový hormon - vazebný protein	4,5, 65, 67, 85, 86, 88, 89, 126, 127, 128, 129, 130, 132, 133, 134, 135

Tab. 21: Porovnání metod

	Protein				
Metoda	**HGH**	**HGHbp**	**Barnase**	**Barstar**	**IM9**
ASEdb	172, 175, 176, 178	43, 104, 105, 165, 169	27, 58, 59, 73, 87, 102	29, 35, 39	33, 34, 41, 50, 51, 55
Robetta-Ala	18, 25, 42, 45, 46, 64, 168, 171, 175, 179	43, 76, 104, 127, 169	27, 59, 60, 87, 102	29, 35, 39, 42, 76	30, 33, 38, 50, 55
S-Transformace z [3]	18, 25, 42, 47, 65, 168, 172, 175, 178, 180	43, 103, 105, 127, 165, 170	27, 58, 60, 73, 87, 102	29, 35, 38, 42	33, 41, 50, 51, 55
S-Transformace	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 25, 26, 27, 28	4,5, 65, 67, 85, 86, 88, 89, 126, 127, 128, 129, 130, 132, 133, 134, 135	6, 7, 43, 44, 54, 56, 57, 58, 59, 60, 86, 87, 90, 91, 92, 93, 101, 103, 104, 105	8, 9, 10, 11, 12, 14, 15, 54, 55, 56, 67, 69	8, 9, 10, 11, 15, 17, 26, 27, 38, 40, 48, 50, 51, 60, 61, 62, 63, 65

V této tabulce HGH – lidský růstový hormon, HGHbp – receptor lidského růstového hormonu, Barnase – endonukléza z bacillus amyloliquefaciens (Barnase), Bastar – endonukleáza z bacillus amyloliquefaciens, IM9 - Colicin E9 imunitní protein z Escherichia

coli. Jak je vidět má metoda produkuje daleko více míst označených jako aktivní místa než původní metoda využívající S-Transformace [3].

Tab. 22: Výkon metody na jednotlivých proteinech

	Parametr				
Protein	**Sn[%]**	**Sp[%]**	**PPH[%]**	**NPH[%]**	**PMU[%]**
Fibroblásotvý růstový faktor	0	88,73	0	96,94	86,30
Endonukleáza C	0	75,84	0	97,41	74,34
TRP RNA-Vazebný útlomový protein	0	81,69	0	93,55	77,33
Lidský alpha hemoglobin	0	89,71	0	96,06	86,52
Lidský růstový hormon	8,33	92,70	4,17	93,75	87,37
Endonukleáza (Barstar)	0	85,71	0	93,51	80,90
Endonukleáza (Barnase)	57,14	84,47	20	96,67	82,73
Interleukin – 4	0	84,25	0	98,17	82,95
Colicin E9 imunitní protein	37,50	80,77	16,67	92,65	76,74
Lidský růstový hormon - vazebný protein	14,29	91,92	5,88	96,81	89,27

V této tabulce ***Sn*** představuje sensitivitu. Ta se vypočítá dle vzorce $Sn = \frac{Pp}{Pp + Fn}$, kde ***Pp*** jsou pravdivě pozitivní, ***Fn*** jsaou falešně negativní. ***Sp*** označuje specificitu. Ta je počítána dle vzorce $Sp = \frac{Pn}{Fp + Pn}$, kde ***Pn*** jsou pravdivě negativní a ***Fp*** jsou falešně pozitivní. ***PPH*** je pozitivní prediktivní hodnota, počítána dle $PPH = \frac{Pp}{Pp + Fn}$. ***NPH*** představuje negativní prediktivní hodnotu vypočítanou ze vzorce $NPH = \frac{Pn}{Pn + Fn}$. A ***PMU*** je průměrná míra úspěšnosti. Získaná ze vzorce $PMU = \frac{Pp + Pn}{Pp + Fp + Pn + Pn}$. Časté nulové hodnoty u parametru Sn a PPH jsou způsobeny vždy tím, že nebyla ani jedna pozice označena jako pravdivě pozitivní. Tedy metoda nemá dobrou sensitivitu a pozitivní prediktivní metodu (ani v případech kdy zmíněné dva ukazatele nejsou nulové). Což naznačuje, že metoda má problém s rozpoznáním aktivních míst a produkuje řadu míst, která jsou označena jako falešně pozitivní.

Jak je vidět v tabulce č. 23 původní metoda má daleko větší sensitivitu a pozitivní prediktivní hodnotu. Má metoda je lepší v parametrech specificity, negativní prediktivní hodnoty i průměrné míry úspěšnosti. Tedy dalo by se říci, že má metoda je lepší na označení míst která aktivními místy nejsou.

Tab. 23: Porovnání výkonu metod

	Parametr				
Metoda	**Sn[%]**	**Sp[%]**	**PPH[%]**	**NPH[%]**	**PMU[%]**
Má S-Transformace	15,79	86,49	5	95,8	83,4
Původní S-Transformace z [3]	79	59	52	84	67

V tabulce č. 24 jsou ještě jednou uvedeny parametry výkonu metody pro jednotlivé zkoumané proteiny. Ale na rozdíl od tabulky č. 22 je zde jako referenční zdroj vzata původní metoda S-Transformace z [3], místo ASEdb spolu s Robetta – Ala.

Tab. 24: Výkon metody na jednotlivých proteinech

	Parametr				
Protein	**Sn[%]**	**Sp[%]**	**PPH[%]**	**NPH[%]**	**PMU[%]**
Fibroblásotvý růstový faktor	6,67	88,55	6,25	89,23	80,14
Endonukleáza C	0,00	75,68	0,00	96,55	73,68
TRP RNA-Vazebný útlomový protein	20,00	80,00	13,33	86,67	72,00
Lidský alpha hemoglobin	0,00	89,71	0,00	96,06	86,52
Lidský růstový hormon	3,57	91,98	7,14	84,66	78,95
Endonukleáza (Barstar)	15,38	86,84	16,67	85,71	76,40
Endonukleáza (Barnase)	26,32	83,52	25,00	84,44	73,64
Interleukin – 4	15,00	84,40	15,00	84,40	73,64
Colicin E9 imunitní protein	27,27	80,00	16,67	88,24	73,26
Lidský růstový hormon - vazebný protein	9,76	92,07	23,53	80,32	75,61
Suma	12,65	86,08	11,54	87,29	76,87

Jak je vidět v tabulce 24 hodnota parametru sensitivity se při porovnávání s původní prací mírně sníží. Naopak hodnoty pozitivní prediktivní hodnoty se více jak zdvojnásobí. Zároveň mírně poklesnou ukazatele NPH a PMU.

V tabulce 25 je vidět porovnání výkonů dostupných metod s mojí S-Transformací (jsou využity parametry vypočítané vůči ASEdb + Robetta – Ala).

Tab. 25: Porovnání výkonu dostupných metod

	Parametr				
Metoda	**Sn[%]**	**Sp[%]**	**PPH[%]**	**NPH[%]**	**PMU[%]**
Má S-Transformace	15,79	86,49	5	95,8	83,4
Původní S-Transformace z [3]	79	59	52	84	67
Digitální filtrace	79,17	79,75	54,29	92,65	79,61
KFC Server	79,17	83,5	59,58	92,96	82,52
Hotsprint	58,33	86,08	56	87,18	79,61
HotPOINT	58,33	81,01	48,28	86,49	75,73
ISIS	33,42	67	32,53	76,81	59,22

7. Závěr

Cílem této práce bylo seznámit s aktivními místy a metodami jejich detekce. A vytvoření programu založeném na jedné popsané metodě. Praktická část se skládá z návrhu a implantace vybrané metody do prostředí Matlab.

Práce je rozdělena do několika větších teoretických částí (teoretický úvod, metody vyhledávání proteinových vazebních míst a metody predikce aktivních míst) a praktických částí (návrh pseudokódu, softwarové řešení a analýza vybraných proteinů). Teoretické části se věnují složení proteinových rozhraní, jakož i teoretickému rozboru jejich vyhledávání. Zde jsou také podrobně vysvětleny dvě metody využívající převod sekvencí do diskrétního signálu. Jedna z těchto metod (metoda využívající S-Transformaci) je v praktické části převedena na pseudokód a následně implantována do prostředí Matlab.

Podařila se implementace do prostředí Matlab i když s podstatně horšími výsledky než uvedené v práci Efficient Localization of Hot Spots in Proteins Using a Novel S-Transform Based Filtering Approach, Sahu S. S. a Panda G. To může být zapříčiněno několika faktory. Hlavním důvodem je nejspíše nepřítomnost filtrace v S-Transformacích zkoumaných signálů v mé práci. Jelikož v původní práci [3] filtr nebyl popsán (bylo pouze uvedeno jeho použití) a odhad užitečné složky a šumu signálu by byl velice obtížný. Dále je zde možnost využití „různých“ proteinů, jelikož ty se vyskytují v několika isoformách. A v původní práci nejsou přesně popsány které isoformy byli použity. A zajisté je také zajímavé, že v původní práci reprezentované vzorce pro inverzní S-Transformaci se ukázaly jako nepoužitelné. Byl jsem tedy nucen využít vzorce z jiné práce [12].

Těmito faktory jsou zajisté mé výsledky ovlivněny. Proto se implementovaná metoda nejeví nijak zajímavě v porovnání s jinými dostupnými metodami. Má sice vysoké parametry *Sp*, *NPH* a *PMU*, ale ty jsou hlavně ovlivněny vysokým počtem pravdivě negativních výsledků. Do budoucna by šla vylepšit použitím zmíněné filtrace (časově frekvenční filtrace).

Byli zpracovány různé proteiny z různých proteinů. S různými výsledky, ale ani pro jeden ze zkoumaných proteinů nepřekročila hodnota parametru *Sp* 60% (nejlépe v tomto ohledu dopadu protein Endonukleázy (Barnase) z Bacillus amyloliquefacien s hodnotou *Sp* 57,14%). Navíc velmi důležitý ukazatel *PPH* nepřekročil hodnotu 20% (opět pro Endonukleázy (Barnase) z Bacillus amyloliquefacien). S tímto proteinem tedy metoda pracuje nejvýkonněji, ale stále hůře než ostatní dostupné metody.

8. Seznam literatury

[1] ÚSTAV LÉKAŘSKÉ CHEMIE A BIOCHEMIE, 2. lékařské fakulty Univerzity Karlovy, Proteiny[online],[cit.2011-12-30], dostupné z: http://www.lf2.cuni.cz/Ustavy/biochemie/vyuka/proteiny1.ppt

[2] NAVRÁTIL,T., MALBOHAN, I. M., Aminokyseliny [online],[cit.2011-12-30], dostupné z: http://biochemie.euweb.cz/Biochemie/Aminokyseliny.ppt

[3] SAHU, S. S., PANDA, G. Efficient Localization of Hot Spots in Proteins Using a Novel S-Transform Based Filtering Approach, IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, no. 5, pp. 1235-1246, 2011

[4] FERNANDEZ-RECIO, J., Prediction of protein binding sites and hot spots. WIREs Computational Molecular Science, vol. 1, pp. 680-698, 2011

[5] NGUYEN, Q.T., FABLET, R., PASTOR, D.,Protein Interaction Hotspot Identification Using Sequence-based Frequency-derived Features, IEEE Trans Biomed Eng. 2011 July 7. [Epub ahead of print] PubMed PMID: 21742567

[6] T. Kortemme, D. E. Kim, and D. Baker, "Computational Alanine Scanning of Protein-Protein Interfaces," Sci. STKE, vol. 2004, no. 219,pp. pl2–, 2004.

[7] WINDOW FUNCTION, poslední aktualizace 23.12.2011 v 16:17 [cit.2011-12-30], Wikipedie, dostupné z: http://en.wikipedia.org/wiki/Window_function

[8] N. Tuncbag, O. Keskin, and A. Gursoy, "HotPoint: hot spot prediction server for protein interfaces," Nucleic Acids Research, vol. 38, no. suppl2, pp. W402–W406, 2010.

[9] V. Veljkovic, I. Cosic, B. Dimitrijevic, and D. Lalovic, "Is it Possible to Analyze DNA and Protein Sequences by the Methods of Digital Signal Processing?," IEEE Trans. Biomedical Eng., vol. BME-32,no. 5, pp. 337-341, May 1985.

[10] KŘÍŽ, Z., NÁRODNÍ CENTRUM PRO VÝZKUM BIOMOLEKUL, PŘÍRODOVĚDECKÁ FAKLUTA MU, BRNO, Docking (dokování) [online],[cit.2012-1-18], dostupné z: http://www.ncbr.muni.cz/school06/downloads/Docking-Kriz.pdf

[11] HOŘEJŠÍ, V., Poněkud neobvyklé membránové proteiny, Vesmír, Vesmír 74, 625, 1995/11 [online],[cit.2012-1-18], dostupné z: http://vesmir.cz/clanek/ponekud-neobvykle-membranove-proteiny

[12] STOCKWELL, R. G., Why use S-Transform?, Fields Institute Communications, vol. 52, 2007, s. 279-310, ISBN 978-0-8218-4276-8

9. Přílohy

P1 – Obrázek 24: Amplitudové spektrum S-Transformace lidského fibroblásotvého růstového faktoru

P2 – Obrázek 32: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem pro endonukleázu C z Cellulomonas fimi

P3 – Obrázek 39: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem pro TRP RNA-Vazebný útlumový protein z Bacillus subtilis

P4 – Obrázek 46: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem pro lidský alpha hemoglobin

P5 – Obrázek 53: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem pro lidský růstový hormon

P6 – Obrázek 60: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem pro Endonukleáza z Bacillus amyloliquefaciens (barstar)

P7 – Obrázek 67: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem pro Endonukleáza z Bacillus amyloliquefaciens (barnase)

P8 – Obrázek 74: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem pro lidský Interleukin - 4

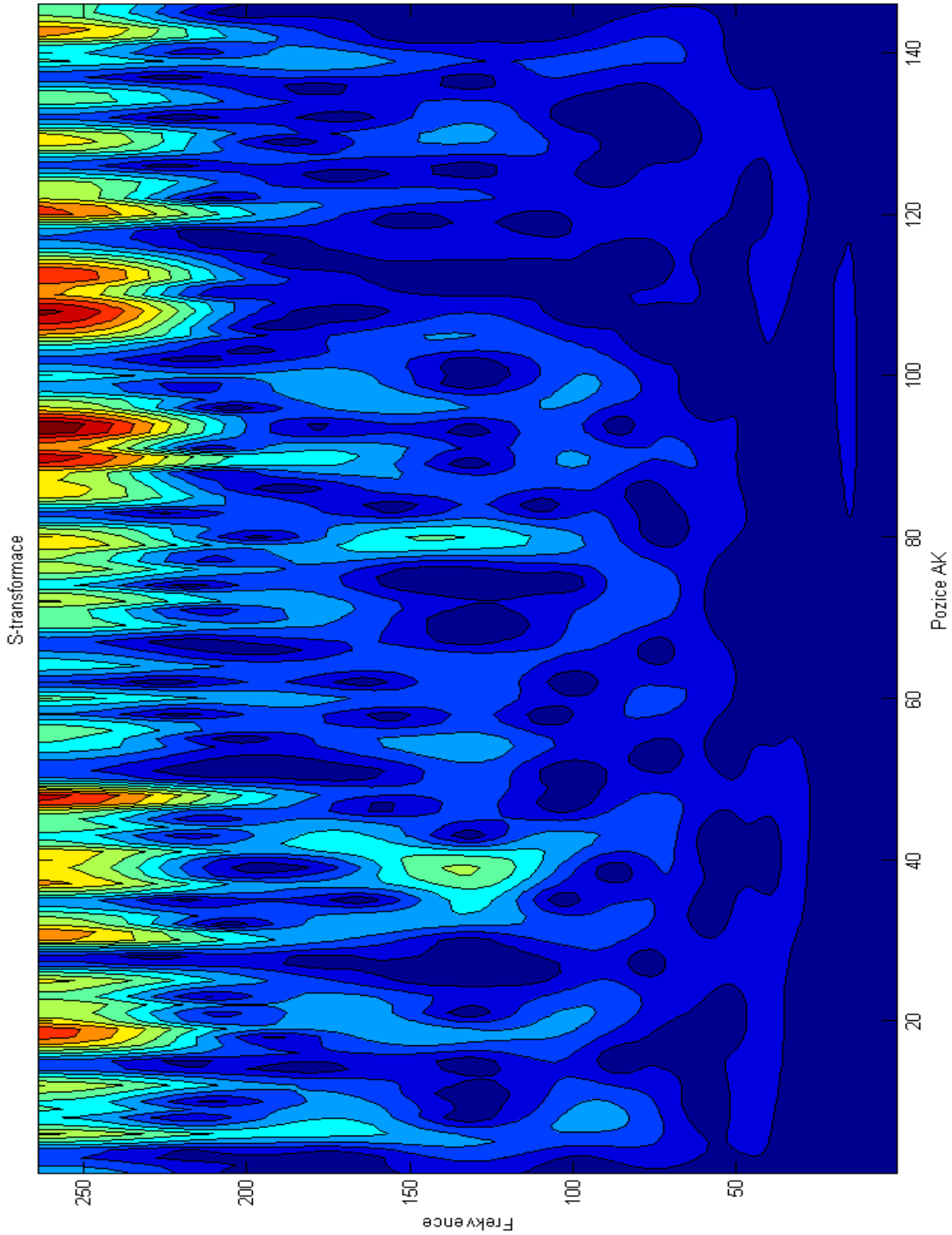
P9 – Obrázek 81: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem pro Colicin E9 imunitní protein z Escherichia coli

P10 – Obrázek 88: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem pro receptor lidského růstového hormonu

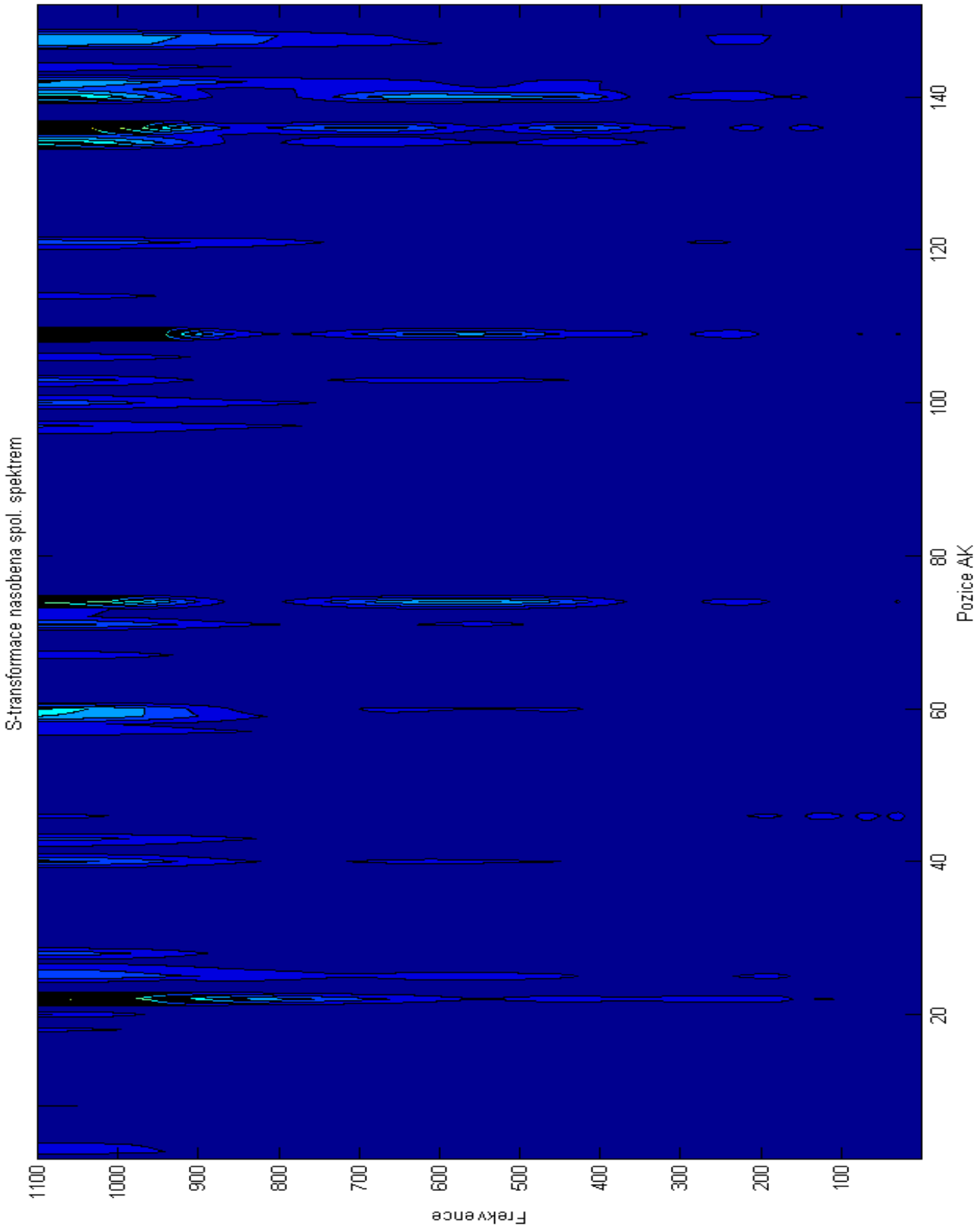
P11 – Vytvořený program se nachází na přiloženém CD v adresáři „program“

P12 – Použité sekvence se nachází na přiloženém CD v adresáři „program\sekvence“

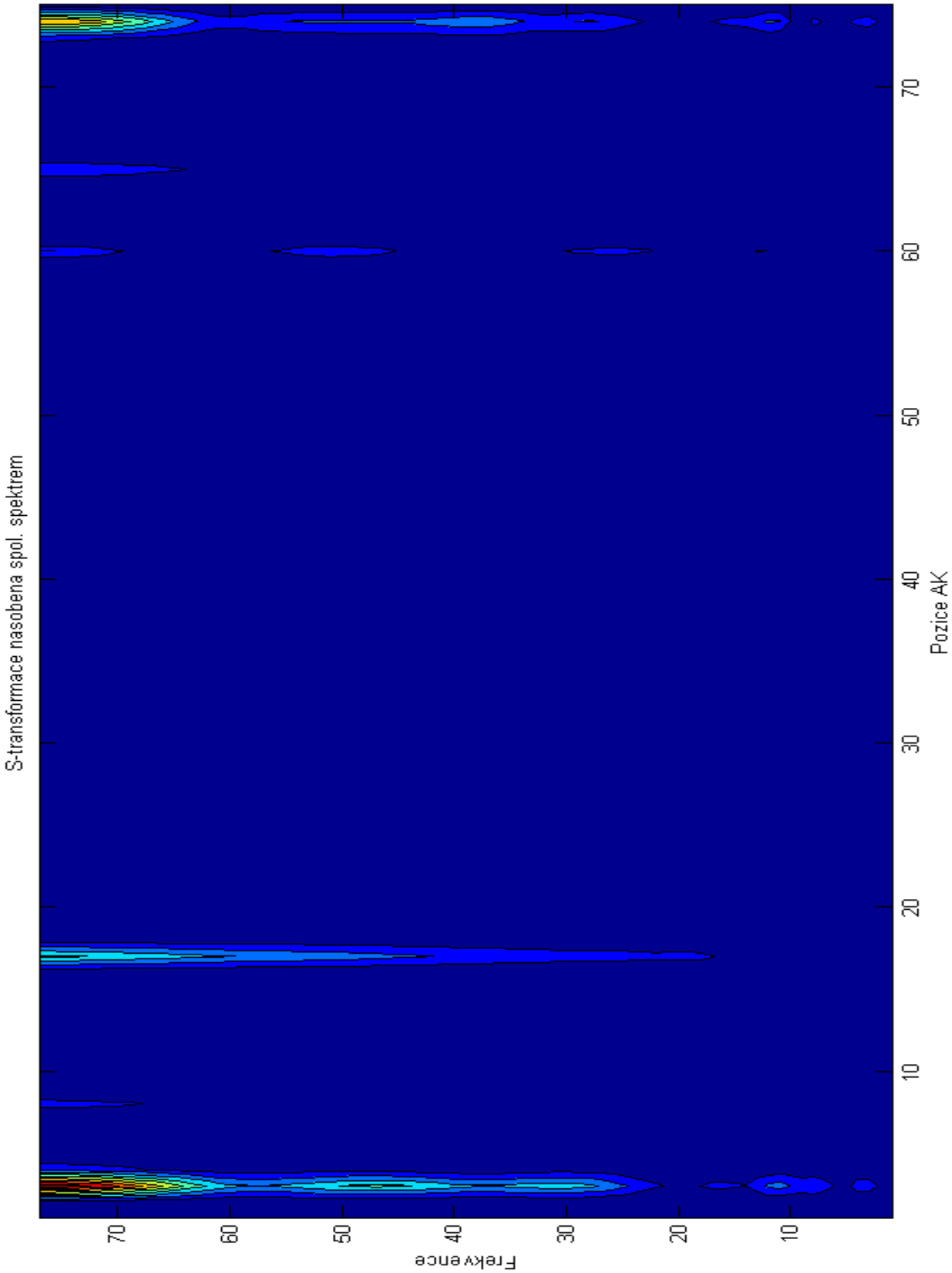
P1 – Obrázek 24: Amplitudové spektrum S-Transformace lidského fibroblásotvého růstového faktoru



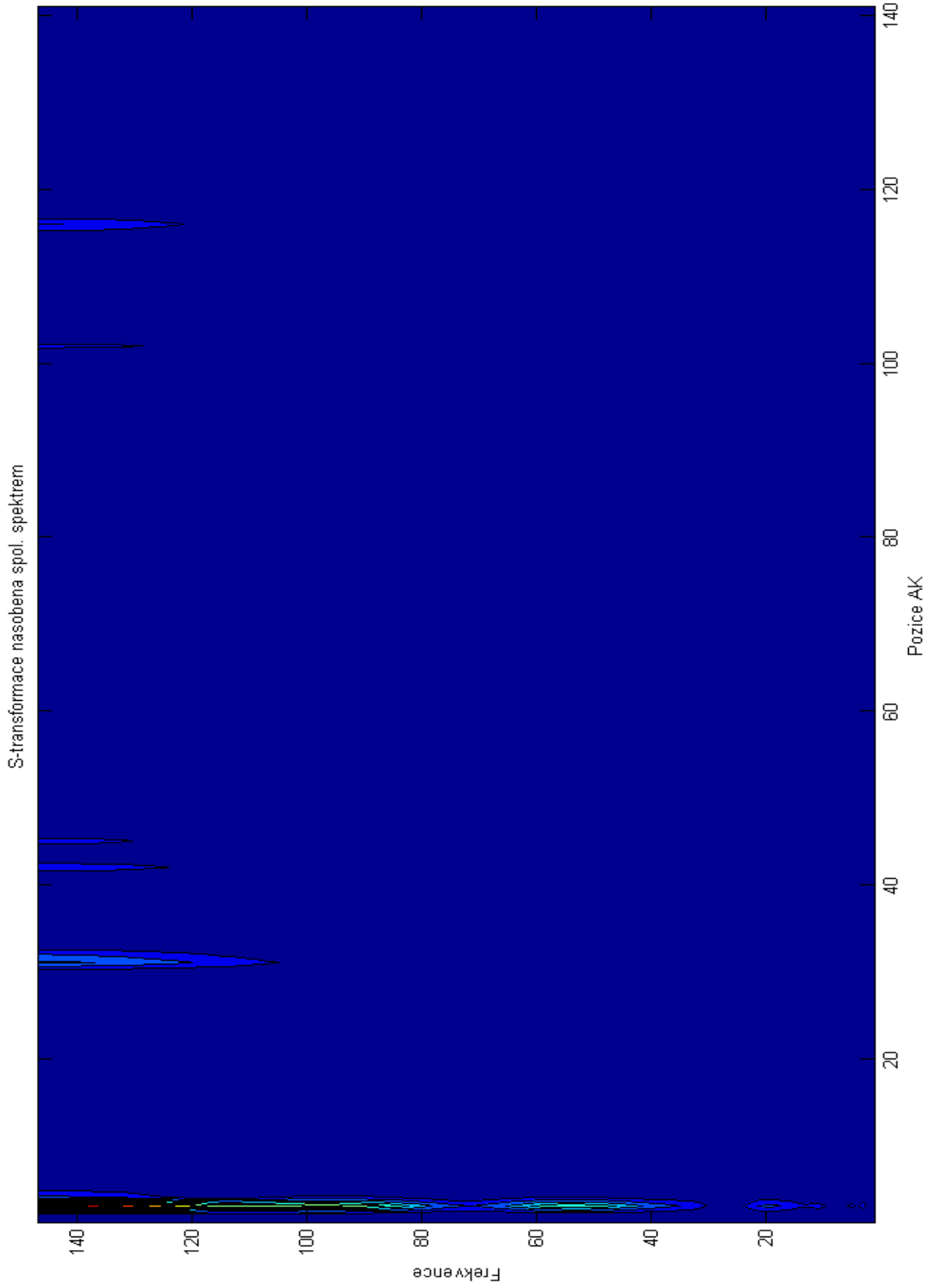
P2 – Obrázek 32: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem pro endonukleázu C z Cellulomonas fimi



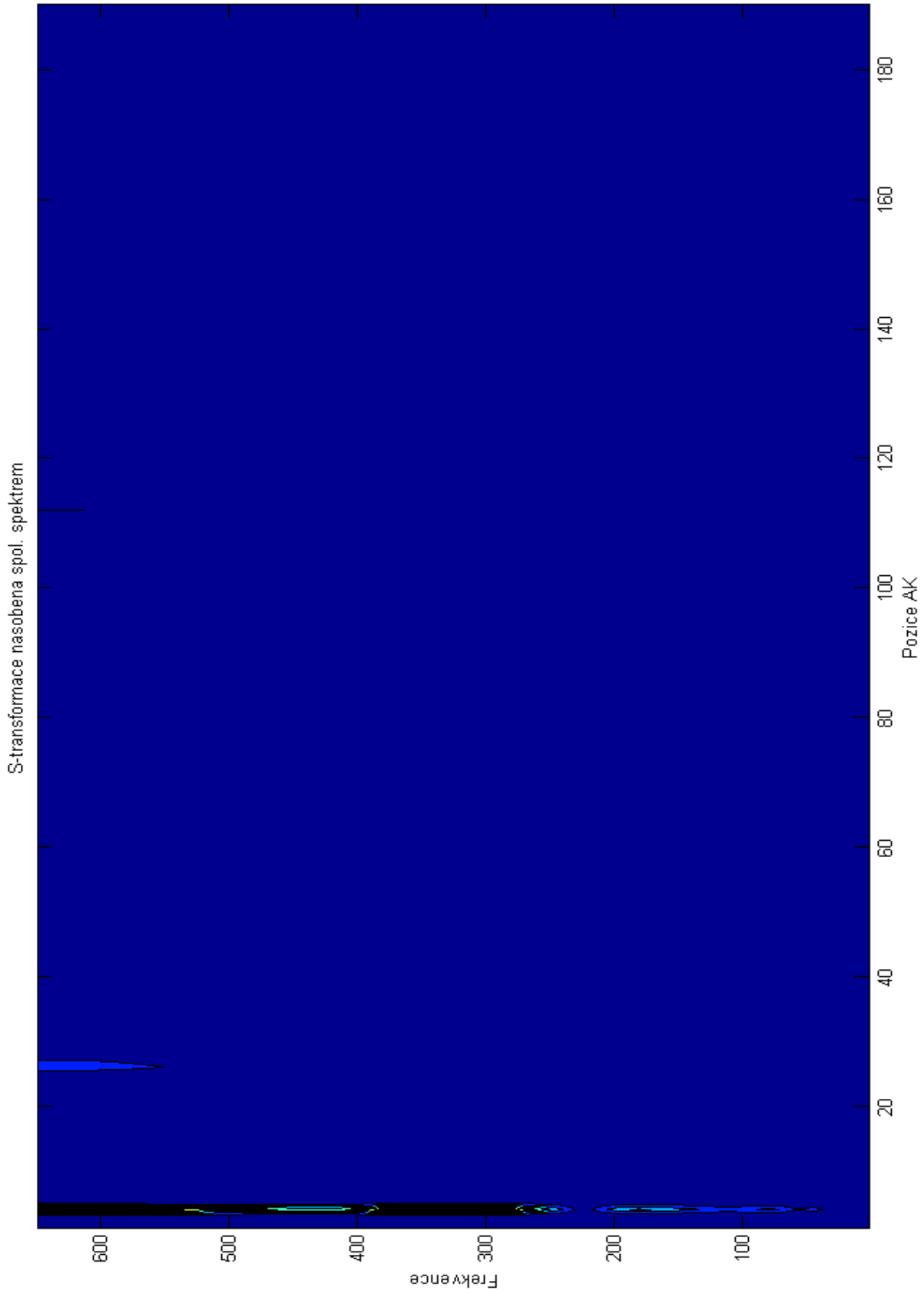
P3 – Obrázek 39: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem pro TRP RNA-Vazebný útlumový protein z Bacillus subtilit



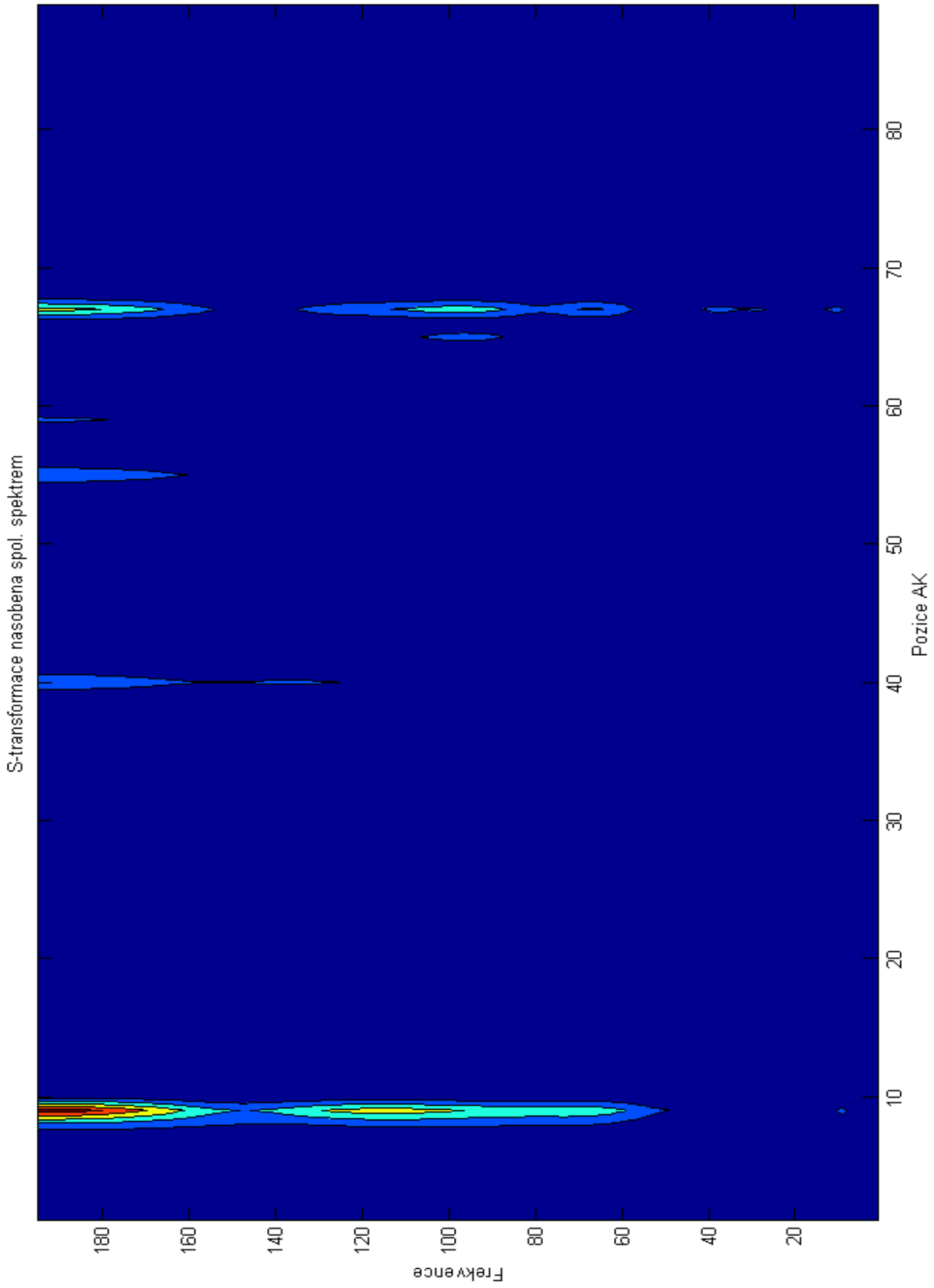
P4 – Obrázek 46: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem pro lidský alpha hemoglobin



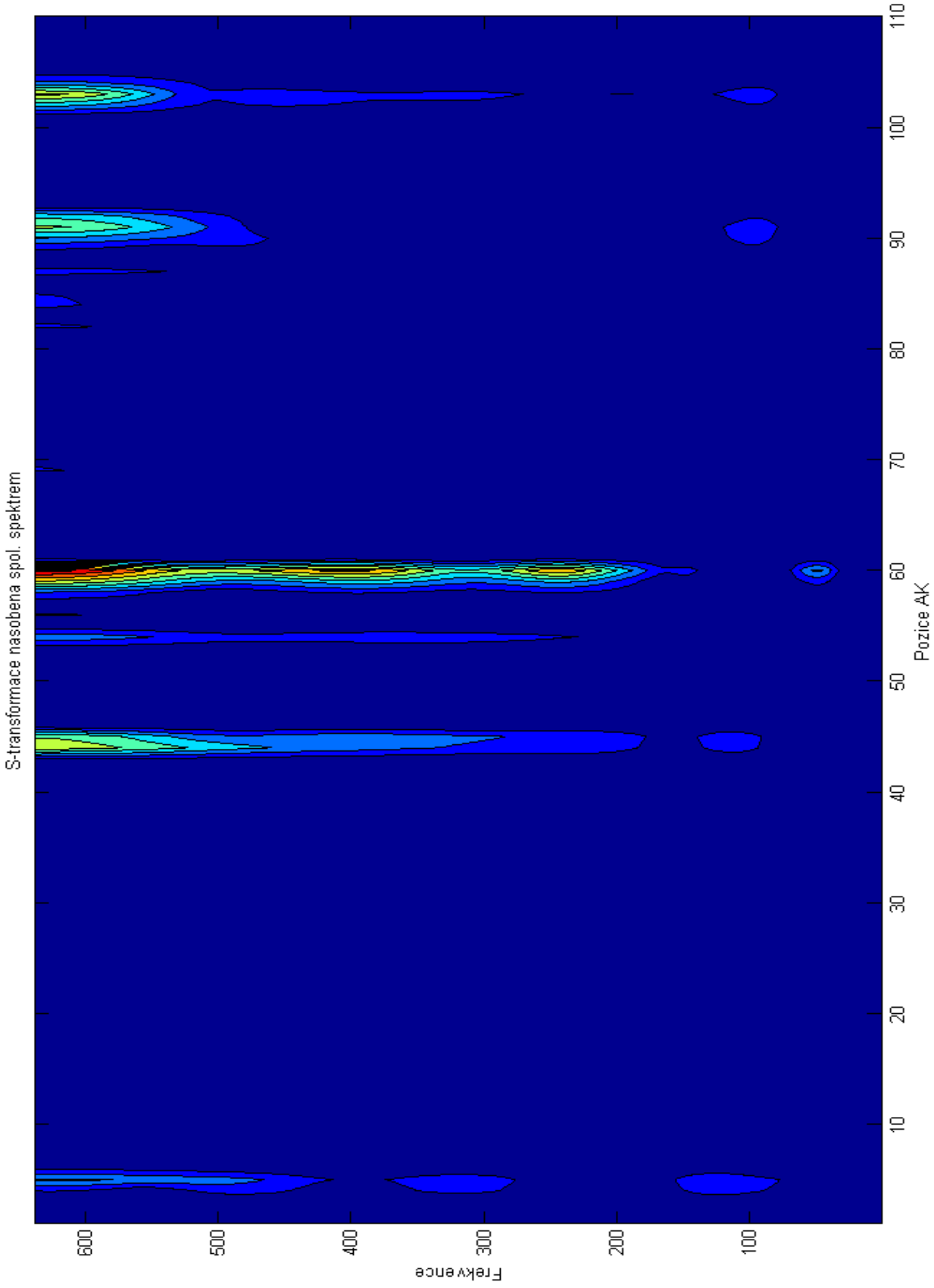
P5 – Obrázek 53: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem pro lidský růstový hormon



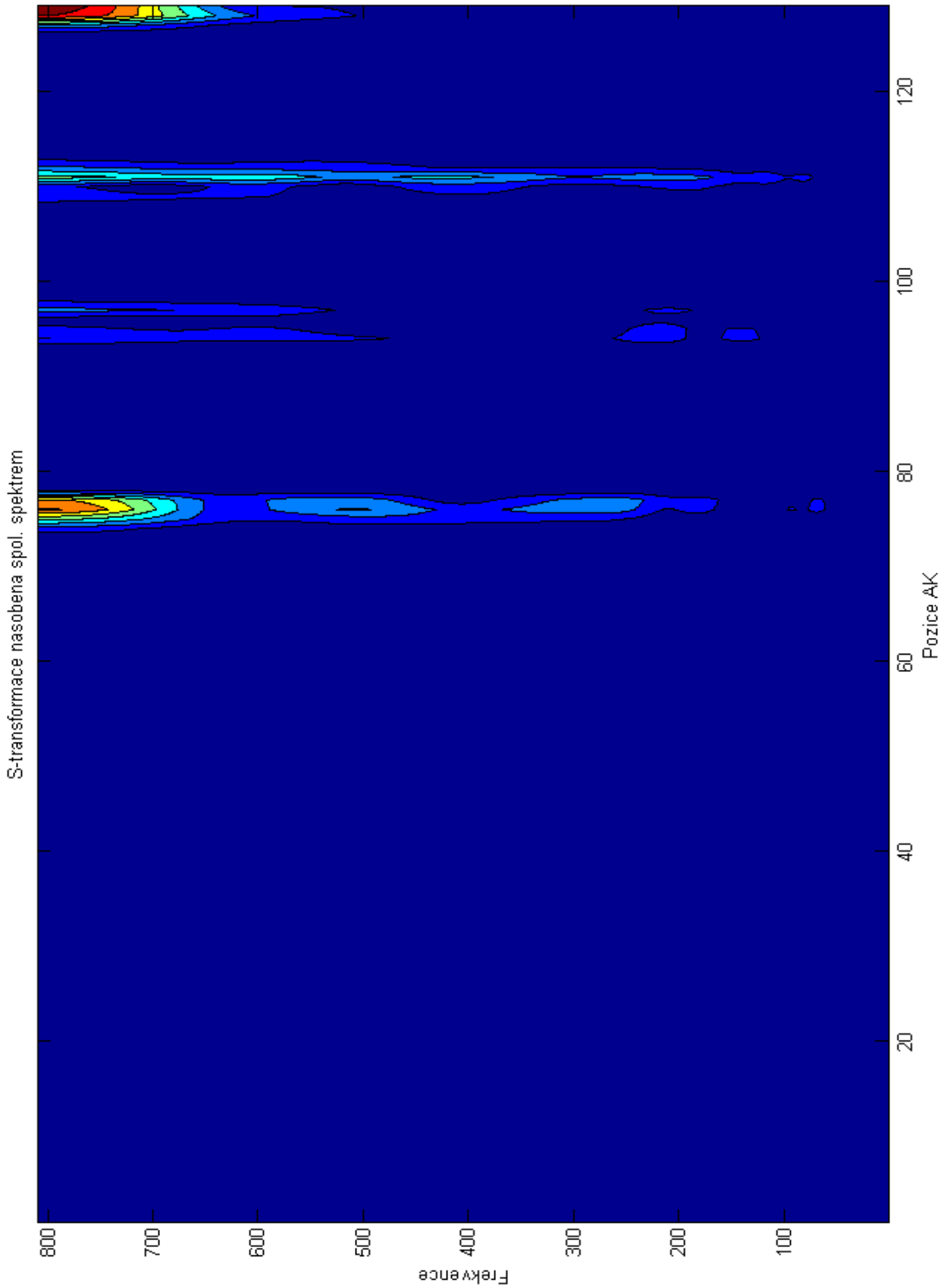
P6 – Obrázek 60: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem pro Endonukleáza z Bacillus amyloliquefaciens (barstar)



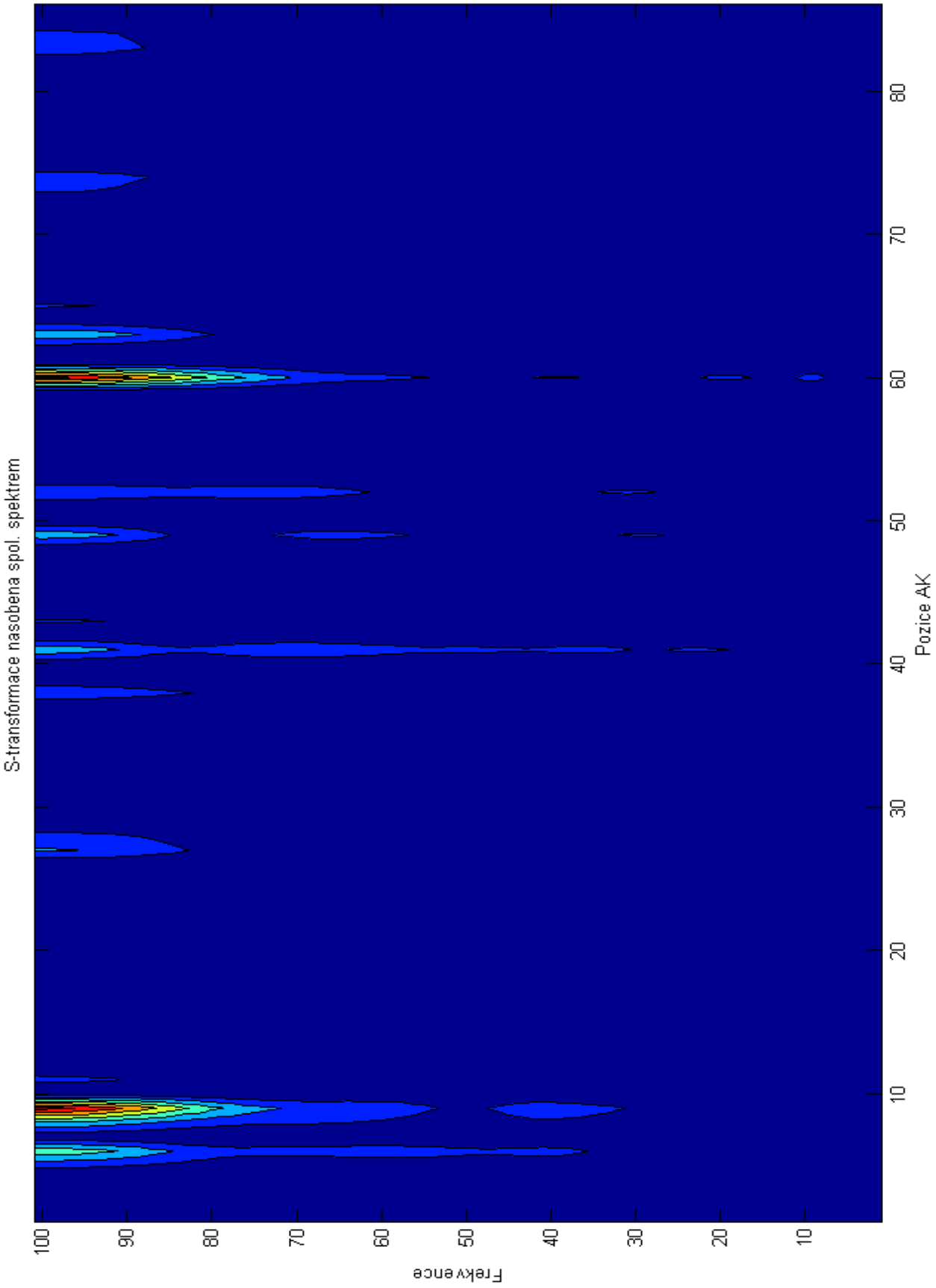
P7 – Obrázek 67: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem pro Endonukleáza z Bacillus amyloliquefaciens (barnase)



P8 – Obrázek 74: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem pro lidský Interleukin - 4



P9 – Obrázek 81: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem pro Colicin E9 imunitní protein z Escherichia coli



P10 – Obrázek 88: Amplitudové spektrum S-Transformace po vynásobení se společným spektrem pro receptor lidského růstového hormonu